2014

A validity argument for the use of scores from a web-search-permitted and web-source-based integrated writing test

Hee Sung Jun
Iowa State University

A validity argument for the use of scores from a web-search-permitted and web-source-based integrated writing test

by

Hee Sung Jun

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol Chapelle, Major Professor
Volker Hegelheimer
John Levis
Frederick Lorenz
Gregory Wilson

Iowa State University

Ames, Iowa

2014

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

While working on this dissertation project, I thanked God repeatedly for how blessed I was to have such generous people around me. It was only with their kind help that I was finally able to finish this dissertation.

I would first like to express my sincere and deep gratitude to my committee chair and major professor, Dr. Carol Chapelle, and my committee members, Dr. Volker Hegelheimer, Dr. John Levis, Dr. Fred Lorenz, and Dr. Greg Wilson, for their guidance and support throughout the course of this project. I have gained so much new knowledge and understanding in many aspects of applied linguistics and statistics through their courses, lectures, publications, and conversations. Their insightful comments helped shape my dissertation, while their enthusiasm for their respective fields and genuine interest in students' learning made them amazing professorial role models to look up to.

Furthermore, I would like to thank my colleagues and the department faculty and staff for making my time at Iowa State University a wonderful experience. I was always met with a smile and kind words, and I will dearly miss everyone.

My appreciation also goes out to the students and instructors who willingly participated in my test, surveys, interviews, and/or rating sessions; without their valuable time and help, this dissertation would not have been possible.

Finally, I am eternally grateful to my family and friends for their encouragement and years of patient prayer.

ABSTRACT

The field of language assessment has seen a recent surge of literature on assessment tasks that integrate two or more skills, such as reading and writing. Source-based writing is also gaining much interest in both first and second language studies, with a particular focus on issues relating to source selection and source language use. The purpose of this study is to build a validity argument for the use of scores from a web-search-permitted and web-source-based integrated writing test. Scores from the test are intended to be used as final exam scores in an academic writing course for international undergraduate students at a large research university in the US. The construct that the test is intended to measure is web-researching-to-write or web-source-based writing, which is defined by the course syllabus and teaching/learning activities.

There are seven inferences that make up the validity argument: domain description, evaluation, generalization, explanation, extrapolation, utilization, and implication. This chain of seven inferences connects the target language use domain and observations of performance to scores and leads ultimately to the consequences of test use. Each inference is supported by a warrant, which in turn is supported by one or more assumptions. Each assumption is backed by evidence. Mixed methods were used to collect and analyze data that would become the backing. Data included 48 Camtasia screen capture recordings, 50 test essays, 40 post-test test-taker questionnaire responses, 6 post-test test-taker interviews, 9 follow-up test-taker questionnaire responses, 9 follow-up test-taker interviews, 5 instructor interviews, and documents.

All of the assumptions underlying the seven inferences were at least partially supported by the backing, which means that the overall validity argument can be upheld by the chain of seven inferences. Further research is suggested to produce additional backing in support of the

comparatively weaker inferences. This study contributes to validation research in language assessment by providing an example of a validity argument constructed for low-stakes classroom-based testing. Furthermore, the study introduces the web-search-permitted and web-source-based integrated writing test as a test that has potential to be adopted by various stakeholders and opens up new possibilities for research on integrated language assessment tasks.

CHAPTER 1

INTRODUCTION

Source-based writing or writing from sources is an integral part of academic writing at the post-secondary level, as source-based writing is commonly required of students as course assignments in the higher education context (Wette, 2010). For many of these assignments, students are expected to find relevant sources on a topic and to present a synthesis of findings from the sources or select pieces of information from the materials to support their own arguments. Whereas the sources that student writers refer to used to be mostly printed texts such as books and periodicals in the past, the internet is increasingly becoming the first and main site that student writers turn to for source information (Stapleton, 2005b; Thompson, Morton, & Storch, 2013). Although the internet has the benefit of easy access to a countless number of websites and web documents, writers have to be aware of the fact that anyone can publish online, and thus they need to be judicious when selecting sources. This leads to the importance of internet literacy (Stapleton, 2005a) or more broadly information and communication technology (ICT) literacy (International ICT Literacy Panel, 2002).

As a likely reflection of the widespread use of source-based writing assignments in language and content classrooms, the field of writing assessment is currently moving towards the use of integrated writing tests that provide listening and/or reading texts as source material (Weigle, 2002, p. 67). This trend is particularly marked by the introduction of the integrated task in the writing section of the internet-based Test of English as a Foreign Language (TOEFL) (Chapelle, Enright, & Jamieson, 2008), which is a large-scale high-stakes proficiency test used for admissions decisions by universities in English-speaking countries. Integrated writing tests are also being used in other contexts at the university level (Weigle, 2004), for example, in

placement tests for incoming international undergraduate and graduate students at U.S. universities.

Those who teach academic writing to international and native-speaker American undergraduate students at U.S. universities quickly come to realize the importance of the students' being able to write from sources, particularly web sources. For example, in the syllabus shared across sections of English 101C, an academic writing course for international undergraduate students at Iowa State University, the schedule includes activities that teach selection of credible sources, source use and integration through summarizing, paraphrasing, and quoting, and use of citation styles. Since the four major writing assignments in the course did not require source use in writing, I thought that the final essay writing test at the end of the semester could be utilized as an opportunity for students to display the source-based writing skills that they had practiced in the class. Furthermore, I believe that the scores obtained from such a test would be meaningful and useful for providing information about writers to the English 101C instructors and other stakeholders. The students' performance on the test would indicate whether they had acquired the skills that were taught in the class regarding use of web sources in writing, in addition to the various aspects of basic essay writing, such as organization, development, expression, and correctness.

With this test use in mind, and also with the intention of conceptually expanding on the idea of integrated writing tests by incorporating ideas from internet literacy, I developed a writing test that allowed test takers to search the web for source information during the test-taking session. The test takers were thus not restricted to one or two pre-selected source texts from which to choose information to integrate into their written responses. In my test, I also did not restrict students from referring to other online language help tools, such as dictionaries,

thesauruses, spell checkers, and grammar checkers. As a first step towards validating the use of scores from this web-search-permitted integrated writing test, I piloted the test with a section of English 101C in Spring 2011. The pilot study results showed that the test was feasible for use as a final exam in English 101C and provided ideas about the adequate time limit for the testing session. The study also produced some preliminary findings about students' writing processes, written products, and perceptions of the test.

My research goals for this dissertation study were to collect and analyze further evidence for the interpretation and use of scores from the test as an end-of-semester summative achievement test in English 101C and to present the evidence in a validity argument using an argument-based approach to validation (Kane, 1992, 2006, 2012, 2013). In particular, I was interested in (a) the test-taking processes of the students, (b) the quality of the essays that were written, particularly in terms of source selection and source language integration, and (c) the perceptions of the test takers and other potential test users and stakeholders regarding the test.

The study employed both quantitative and qualitative methods in a mixed methods research design. To investigate what test-taking processes the students go through, screen-capturing software called Camtasia was used to record the computer screens as the students were taking the writing test. The screen recordings were coded for both internet search behaviors and writing behaviors. To examine the quality of essays, discourse analysis was conducted for credibility of sources selected and integration of source language into the essays. Lastly, the perceptions of test takers were investigated through a post-test questionnaire, post-test interviews, a follow-up questionnaire, and follow-up interviews, while those of other potential test users and stakeholders were collected through instructor interviews.

All of the data and the results obtained from the analyses were used as evidence that supported the interpretive argument for the use of scores from the web-search-permitted and web-source-based integrated writing test. The resulting validity argument could inform potential test users of the validity of using scores from the web-search-permitted integrated writing test as final exam scores in undergraduate academic writing courses. Potential users of the test could be writing instructors and/or coordinators of writing programs. Furthermore, the results may contribute to broadening the research possibilities in the field of writing assessment, which has already begun to expand to integrated writing tests. Future research could even address the potential use of a web-search-permitted test for large-scale testing purposes.

The remainder of the dissertation will be organized as follows. A review of the literature that provided the background for the development of the integrated writing test will be presented in Chapter 2, along with the interpretive argument that drove the validation project by providing the framework for the study. The methodology used to collect and analyze the evidence that is needed to support the interpretive argument will be described in Chapter 3, while the results of data analysis will be presented and discussed in Chapter 4. The dissertation will conclude with the validity argument, which combines the interpretive argument and the evidence in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

This chapter first reviews the previous literature in four areas that form the basis of the current study, namely, (a) source-based writing, (b) internet literacy, (c) use of help options during a writing test, and (d) integrated writing tests. This review of the literature is followed by an interpretive argument for the use of scores from the web-search-permitted and web-source-based integrated writing test. The chapter ends with a list of research questions for the current study, which are based on the backing that is needed to support the interpretive argument.

Source-Based Writing

The first main area of research that forms the basis of integrated writing tests and consequently this dissertation study is source-based writing, also termed as writing from sources or content-responsible writing. With regard to the reading-writing connection, Hirvela (2004) discusses both writing-to-read and reading-to-write in connecting reading and writing in second language (L2) writing instruction. Writing-to-read includes summarizing what one has read to understand the reading better, while reading-to-write involves using reading as a source of ideas and support for writing. The focus of this dissertation study is on reading-to-write.

Taking both first language (L1) and L2 writers into consideration, Delaney (2008) takes a constructivist view of literacy tasks in defining the reading-to-write construct as "a reciprocal interaction between literacy skills, in which the basic processes and strategies used for reading and writing are modified by an individual's goals and abilities, and also by external factors" (p. 141) and "a dynamic activity that interacts with task demands and individual factors" (p. 147) such as language proficiency and educational level. Results from Delaney's empirical study

using a summary task and a response essay task suggest that reading-to-write ability involves not only reading for comprehension, but also reading with the goal of selecting information from the source text to help the reader/writer in writing his or her own text. Delaney further suggests that reading-to-write ability is different from writing texts without the use of sources, thereby implying that the reading-to-write construct is a unique ability in itself and not merely a sum of reading ability and writing ability. This view is supported by Gebril and Plakans (2009) who suggest that "a construct for integrated reading-writing includes ability to control and create language elements of the written product, knowledge and use of integrating source text, as well as general L2 proficiency and L2 reading ability" (p. 70).

Source-based writing is considered to consist of multiple skills. In his book chapter on teaching the academic essay using a genre approach, Dudley-Evans (2002) cites Jordan (1997) who noted that academic essay writing requires numerous skills such as (a) planning, writing drafts, revising; (b) summarizing, paraphrasing, and synthesizing; (c) continuous writing in an academic style organized appropriately; (d) using quotations, footnotes, bibliography; and (e) finding and analyzing evidence, using data appropriately. These are general writing skills that apply to many genres that university students need to write, although (b), (d), and (e) are particularly relevant for source-based writing. Dudley-Evans (2002) more specifically discusses the use of sources by suggesting that when teaching students how to transform knowledge by evaluating sources in their academic essays, "a full discussion of all the issues and a detailed examination of the techniques of referencing, both in the text and in the bibliography, are more effective than a focus on the techniques of paraphrasing" (p. 233). He particularly stresses the importance of "teaching the technique of making a reference in the text" (p. 233), while considering what is acceptable in the different disciplines. Dudley-Evans (2002) also cites

Wilson (1997) who suggested that there are four stages in the development of academic writing with regard to the use of sources: (a) repetition, (b) patching, (c) plagiphrasing, and (d) conventional academic writing. Wilson noted that students at the third stage are beginning to speak with their own voices and are on their way to developing an appropriate academic writing style.

While presenting materials for teaching graduate students to write literature reviews, one of the representative genres of source-based writing, Swales and Lindemann (2002) identify macro features, mid-level functions, and micro functions to be taught. Macro features, which are mainly reading activities, include aims and purposes, library searches, and taking notes, while micro functions, which are mainly mechanics of writing, include tense, citation, reporting verbs, and adjuncts of reporting. The mid-level functions, which link the macro and micro levels and represent meaning-making activities, are how to use citations, what and how much to cite, paraphrasing and synthesis, and ways of organizing the literature review.

It is a generally held belief among scholars in the field of writing and composition that source-based writing is an important part of academic writing, particularly at the post-secondary level (e.g., Delaney, 2008; Erling & Richardson, 2010; Li, 2013). In second language studies as well, Weigle (2002) suggests that for writing at the tertiary level, "a strong case can be made for writing both to be based on a reading and to require students to provide appropriate and relevant support for ideas" (p. 95), since writing in the university is almost always "based on some prior reading" (p. 95) and university writing assignments usually require referring to written sources. Textbooks and writer's handbooks that teach source use in writing are plentiful (e.g., Faigley, 2006; Harris, 2011). These books are most often used in composition courses for first-year

college students in the U.S., but they are also used as references in writing courses for more advanced college students and international undergraduate and graduate students.

Furthermore, survey studies reveal the prevalence of source-based writing in university writing and content courses. For example, Burton and Chadwick's (2000) survey of 543 college students found that 94.9% of the students who had written a paper for a class in the past year used outside internet and/or library sources. In Carson's (2001) survey of six content courses at a U.S. university, it was found that undergraduate students in an introductory history class were required to use information from lecture material, handouts, the primary text, a dictionary, and additional source material from the library in writing a take-home essay exam. The students employed numerous cognitive processes and writing skills (e.g., paraphrasing, synthesizing, summarizing, providing appropriate support) while preparing for and producing the essay. Graduate students in biology had to write a critical review of a recent article using the course text, a computer database, related journal articles, and class notes, while those in history wrote a research report based on archival research. The final exams for biology and psychology graduate students also involved synthesizing information from texts.

At Iowa State University, the two first-year composition courses English 150 (Critical Thinking and Communication) and English 250 (Written, Oral, Visual, and Electronic Composition) give students several source-based writing assignments. In English 150, two assignments require source use: a profile of a campus program or organization based on analysis of public documents and a report and commentary on a campus landscape, building, or art. In English 250, students write a summary of a text and an argumentative research paper (Department of English, Iowa State University, 2012).

In light of the popularity of source-based writing assignments, there are several issues that have been problematized and dealt with quite extensively in both first (L1) and second language (L2) writing research, such as source selection (e.g., Thompson, Morton, & Storch, 2013), plagiarism (e.g., Abasi & Graves, 2008; Bloch, 2001; Flowerdew & Li, 2007; Pecorari, 2001), textual borrowing (e.g., Barks & Watts, 2001; Eckel, 2011; Hirvela & Du, 2013; Shi, 2004, 2007; Weigle & Montee, 2011; Weigle & Parker, 2011; Wu, 2013), and correct use of citations and reference lists (e.g., Kim, 2009; Mansourizadeh & Ahmad, 2011; Petrić & Harwood, 2013). These issues are particularly problematic for L2 writers, who need to deal with both linguistic deficiencies and cultural differences, all the while trying to develop "identities as academic writers and members of a disciplinary community" (Wette, 2010, p. 158). For example, in Wu's (2013) study, 40 essays from a reading-to-write test were analyzed to discover that non-native writers used more mechanical copies of the source text than native writers. However, Pecorari (2003) suggests that L2 writers often plagiarize without the intention to do so and points out that proactive teaching is a more desirable way to prevent plagiarism rather than "post facto punishment" (p. 317). Specific suggestions as to exactly what to teach L2 writers regarding source use are given by Davis (2013):

> [T]he amount of citation expected, the different functions and use of integral and non-integral citation, the use of source words, the ways to synthesise, the use of internet sources, the range of reporting verbs and how to progress from patchwriting to effective paraphrasing by improving vocabulary learning. (p. 134)

Fortunately, a growing number of studies are addressing the challenges that novice L1 and L2 writers experience with source use in writing, and research is continuing to be conducted to develop effective ways of teaching students the skills of writing from sources. For example,

one recent action research study by Wette (2010) presents an 8-hour unit that explicitly teaches students the correct use of sources in writing and evaluates student learning in a university-level English for academic purposes (EAP) class. Upon completion of the unit, the 78 students that participated in Wette's study gained declarative knowledge about the "technical and rule-governed aspects of writing from sources" (p. 168), such as deciding what to cite, citation formatting, and quoting. Brown, Dickson, Humphreys, McQuillan, and Smears (2008) introduce a web-based anti-plagiarism unit composed of six multimedia lectures that teach referencing skills and the use of referencing software to undergraduate students. Questionnaires were used to investigate the changes in the perceptions of the students regarding referencing. Niedbala and Fogleman (2010) discuss teaching information literacy using a course wiki.

At the graduate level, Dovey (2010) claims that it is important to focus on the process as well as the product when teaching the genre of literature-based reports to postgraduate EAP students. She goes on to describe a 13-week course in which students complete a series of recursive assessment tasks that involve constructing graphic organizers and concept matrices to "scaffold and facilitate the processes of organi[z]ing, selecting, and integrating information from multiple sources" (p. 58). The course evaluations showed that the students were particularly appreciative of how the teaching and learning process was incremental. A study by Stapleton (2012) demonstrates how plagiarism-detection software can be used in writing classes to deter students from plagiarizing. In his study, Stapleton implemented Turnitin, a popular anti-plagiarm service, in two graduate writing classes and found that the class which was aware of the software being used to evaluate the originality of the essays committed less copying and intentional plagiarism compared to the class which was not aware of Turnitin.

Internet Literacy (and Other New Literacies)

The second idea that is important in understanding the thinking behind the development of the web-search-permitted test is internet literacy. Numerous survey studies have found that college students are now rapidly turning to the internet as the primary source of materials for writing as opposed to printed materials in the past (e.g., Biddix, Chung, & Park, 2011; Burton & Chadwick, 2000; Jones, 2002; Jones, Johnson-Yale, Millermaier, & Perez, 2008). Jones (2002) surveyed 2,054 college students to find out that 73% used the internet as the primary source for research, while only 9% said that they used the library more than the internet when searching for information. In a follow-up survey of 7,421 U.S. college students at 40 campuses, Jones, Johnson-Yale, Millermaier, and Perez (2008) found an increase in the academic use of the internet. Specifically, 95% of students reported using search engines followed by library websites (68%), news websites (64%), online encyclopedias (48%), and other sources (10%) when searching for information online. This popularity of web sources has led to the belief that it is critical for college students to practice and attain the skills of evaluating the credibility and reliability of web sources when searching for sources to use in writing (Biddix, Chung, & Park, 2011). These skills or abilities have been given different labels in the literature, including internet literacy and web literacy.

With the development of information technologies and the coming of the new media age, both L1 and L2 literature have seen terminologies abound for new literacies such as technology literacy (e.g., Hohlfeld, Ritzhaupt, & Barron, 2010), digital literacy (e.g., Zhang, 2003, 2005), electronic literacy (e.g., Shetzer & Warschauer, 2000), ICT literacy (e.g., International ICT Literacy Panel, 2002; Lowe & McAuley, 2000), web literacy (e.g., Sorapure, Inglesby, & Yatchisin, 1998), and internet literacy (e.g., Stapleton, 2005a). While many of these terms

partially overlap with each other in meaning, internet literacy was chosen to be the single representative term for the current study because of its narrower and more specific meaning. Stapleton (2005a) defines internet literacy as "effectively using the web as a research source" (p. 136). He sees internet literacy as being one aspect of electronic literacy—the skills required for reading and writing texts in an online environment (Hirvela, 2004).

The definitions for the other new literacy terms are usually broader. Web literacy was defined as involving "an ability to recognize and assess a wide range of rhetorical situations and an attentiveness to the information conveyed in a source's nontextual features" (Sorapure et al., 1998, p. 410). ICT literacy was first defined by a panel convened by Educational Testing Service (ETS) as "using digital technology, communications tools, and/or networks to access, manage, integrate, evaluate, and create information in order to function in a knowledge society" (International ICT Literacy Panel, 2002, p. 2). This definition has been expanded since 2002 to become the ICT literacy framework which includes "seven key performance areas: defining a need for information, accessing information via technology, evaluating online information, managing digital information, integrating information from varied digital sources, creating information, and communicating information through technology" (Ali & Katz, 2010, p. 5). Each of these seven areas is accompanied by a list of concrete behaviors that further specify what an ICT literate person can do. Ali and Katz (2010) further explain that "ICT literacy bridges the definitions of information literacy and technology literacy. By focusing on information skills in the context of technology, ICT literacy addresses the key challenges of information access, information overload, and information quality" (p. 5). ETS has also recently developed a test called *iSkills™ Assessments* to measure ICT literacy (Katz, 2007; Katz et al., 2008; Tannenbaum & Katz, 2008).

Because the realm of the internet is virtually limitless and the amount of information that students are exposed to is unimaginably vast, it becomes important for writing teachers to teach students the skills to evaluate the quality of web sources before selecting information from them. Numerous studies have highlighted the potential dangers of using internet sources for academic writing while suggesting pedagogical interventions that may point students in the right direction. For example, Sorapure, Inglesby, and Yatchisin (1998) stress the importance of developing web literacy in student researchers in order to overcome "two major challenges posed by the Web as an information resource—its diverse and unfiltered content and its hypermedia format" (p. 412). They believe that "a close and careful reading of Web sites can enhance students' research and writing skills" (p. 423).

Furthermore, Stapleton and his colleagues have published a series of studies that emphasize the importance of developing internet literacy. Stapleton (2003) points out the bias that may exist in web-based sources and calls for "new initiatives in the EAP writing curriculum" to teach students to recognize those "subtleties of persuasion" (p. 242). Stapleton, Helms-Park, and Radia (2006) analyzed 68 web sources that were chosen by 19 English as a second language (ESL) student writers in research papers and found that 33 of them were "unconventional." These include interest groups, commercial companies, and informal materials. Based on this finding, the authors claim that second language students need to be taught website evaluation skills in academic writing. Helms-Park, Radia, and Stapleton (2007) found that the search engine Google "often steers students towards sites that are not suitable for academic purposes" (p. 70) and also adds confusion for L2 writers who are bombarded with the unlimited possibilities of information, although Google Scholar is just as reliable and effective a source as the library e-searches for finding research material for English for academic purposes students.

Stapleton (2005a) points out that search engines can produce search results that are "skewed towards certain agendas" (p. 138), which can be especially problematic for L2 learners less familiar with the language and perhaps "culturally disadvantaged with regard to recognizing the quality and bias of web-based sources" (p. 141). The author then proposes a four-step plan for teaching evaluation of web sources as a way to develop internet literacy: (a) search engine practice; (b) assessing websites; (c) describing the various forms of weak reasoning associated with bias; and (d) website evaluation assignment. He found that seven students who were trained according to this plan were able to detect bias in websites and even proceed to find better, more balanced sources on the web.

In another study, Stapleton (2005b) analyzed the 243 web sources that 43 Japanese undergraduate English as a foreign language (EFL) students referred to in an essay assignment and investigated through a questionnaire the language-related strategies that students employed in writing their essays using the web. He found that a considerable portion of the web sources chosen by the students (38%) was of questionable quality, and several students even succumbed to the temptation of plagiarizing. Some important teaching implications from this study are (a) to inform students of web genres that are more suitable for use in academic writing; (b) to teach students how to evaluate web sources for their reliability and appropriateness; and (c) to warn students about the unethical use of translators and plagiarizing.

Other pedagogical efforts toward promoting effective web-based research include Helms-Park and Stapleton (2006), who developed a rating instrument for assessing the suitability of web sources based on surveys of faculty members, Wiley et al. (2009), who show the effectiveness of a simple instructional unit that teaches undergraduate students how to evaluate the trustworthiness of information on web sites, and Corbett (2010), who introduces a method that

uses library and internet search tools such as Google to teach library research to students in first-year composition courses.

Use of Help Options during a Writing Test

Although there is considerable literature on the use of language help options and technological tools while writing in general (e.g., Bloch, 2009; Christianson, 1997; Gilmore, 2009; Harvey & Yuill, 1997; Kennedy & Miceli, 2010), research on the use of such help options while writing specifically in a test situation has been relatively scarce. Weigle and Jensen (1997) describe a university-level content-based final exam that was specifically designed to add authenticity and interactiveness, which are two of the six qualities of test usefulness in Bachman and Palmer (1996). The writing portion of the test allows dictionary use, which is seen by the authors as adding authenticity to the test because students would use dictionaries when working on their writing assignments for a content course. Weigle (2002) suggests that "a broader definition of writing ability, in which one uses all available resources, does not necessarily preclude the use of dictionaries" (p. 106). However, she points out that little evidence is available on "the efficacy of dictionary use during a language test" (p. 106). She cites two studies that have found no significant effects of dictionary use on scores on L2 reading tests, although there is an effect on time taken to complete the test (Bensoussan et al., 1981; Nesi & Meara, 1991). Weigle calls for further research in this area. Furthermore, the use of other offline and online help options such as thesauruses, Criterion®, grammar checkers, translators, and corpora during a writing test has not been researched widely.

One notable study that responded to Weigle's call for further research on the issue of dictionary use during a writing test is East (2006), who found that the use of a bilingual

dictionary resulted in increased lexical sophistication, but at the same time, test takers also frequently misused the dictionary look-ups. In addition, using the bilingual dictionary did not result in the improvement of test scores compared to test scores obtained from the no-dictionary condition. However, East suggests that test takers should be provided with training on how to correctly use dictionaries in the test preparation period, as this could result in more effective use of dictionaries during the test, and the dictionary could perform its function as a truly helpful help option that improves the lexical sophistication of test takers' essays.

Integrated Writing Tests

The last major area that forms the background of the current study is integrated writing tests—tests that assess writing in conjunction with other skills such as listening and/or reading. There are different task types that have been developed for use in integrated writing tests. First of all, Yu (2008, 2009, 2013) looked at a summarization task in which test takers read an extended text and summarize the main points of what they have read. A listen, read, and summarize task is currently being employed by ETS in the internet-based TOEFL where a reading text and a listening passage on the same topic but with a different point of view are presented for processing by test takers, and test takers then write a summary in which they compare and contrast the different opinions on the topic. Delaney (2008) used a summary task and a response essay task in her study on the construct of reading-to-write. Gebril and Plakans (2009) discuss a task in which test takers read a text and write an argumentative essay based on the text. This type of task is also discussed by Weigle (2004). As a last example, the English Placement Test at the University of Illinois at Urbana-Champaign uses an integrated task in which test takers listen to a short lecture, read a text, and write an argumentative essay in which they incorporate the two

sources. In short, there is variation in what type(s) of text are provided as input (listening and/or reading) and what type of text test takers are expected to produce (summary or essay).

The rapidly accumulating body of research on integrated writing tests particularly in the past decade signifies the centrality of the topic. At the same time, the widespread use of integrated writing tests is marked by the addition of integrated tasks in the writing sections of large-scale English language tests, such as the internet-based TOEFL, International English Language Testing System (IELTS), and Pearson Test of English (PTE) Academic, as well as in English placement tests at numerous U.S. universities (e.g., Lee & Anderson, 2007; Plakans, 2008).

The use of integrated writing tests is promoted by researchers who posit that integrated writing tests may have more authenticity than prompt-based independent writing tests because they more closely resemble tasks required of students in university courses (Cumming et al., 2005; Gebril & Plakans, 2009; Plakans, 2008). Weigle and Jensen (1997) claim that a writing test that allows students to gain and synthesize information on a topic from numerous sources and use this information to support their point of view is a more accurate reflection of the kind of writing expected of university students. Weigle (2002) sees this as building "authenticity and interactiveness into timed writing tasks" (p. 186).

Weigle (2004) describes an integrated test for non-native speakers, which was developed for university writing examination requirements of the state of Georgia, and argues that the integrated test "can provide more valid information about students' ability to write at the college level than a prompt-based essay test" (p. 29). The support for the use of integrated tests over prompt-based tests comes from (a) numerous research findings that show academic writing to be done almost always as a response to source texts rather than in isolation; and (b) the argument

that source texts allow test takers access to background knowledge and act as stimulus for their own ideas. Furthermore, Weigle's own study results suggest that the integrated test increases rater reliability and consistency across topics while also having positive washback by shifting the focus of the test preparation course from the five-paragraph essay format to skills more applicable to other content courses such as argumentation and appropriate source use.

Many studies have been conducted on the processes of integrated writing tasks and tests. Ruiz-Funes' (1999) case study of a Spanish as a foreign language student's reading-to-write processes in an assignment is one such study, as is Esmaeili's (2002) study of ESL students' self-reported use of writing strategies in a thematically related reading and writing test. Asención (2004), on the other hand, looked at the cognitive processing of second language learners during reading-to-write assessment tasks, while Yang (2009) and Yang and Plakans (2012) investigated the effects of test takers' self-reported use of strategies in a reading-listening-writing test (integrated task on the TOEFL iBT) through both structural equation modeling and qualitative approaches. The latter two studies found that self-regulatory strategies play a large role because the task requires test takers to manage reading, listening, and writing skills within one task. The use of discourse synthesis strategies also had a direct and positive effect on scores.

The processes of integrated writing tests were further researched in a series of four studies by Plakans, who used think-aloud protocol and interview data to compare the composing processes in writing-only and reading-to-write tests (Plakans, 2008), analyze the discourse synthesis process in the integrated writing test (Plakans, 2009a), investigate the reading strategies involved during the integrated test (Plakans, 2009b), and compare test takers' task representation in the two tests (Plakans, 2010). Similar studies followed, including Chan (2011) who used screen video and stimulated recall to investigate the cognitive processes that ten test takers

engaged in during a summarizing test. She found that most test takers engaged in macro-planning before writing and discourse synthesis while writing. Wolfersberger (2013), based on his ethnographic study of four L2 writers completing a source-based argumentative essay assignment, showed how the writers' task representation affected their performance and argued that task presentation should be an important consideration in construct definition for integrated writing assessment.

Investigations of the products were also added to the research base on integrated writing tests. Watanabe (2001) studied the text features of products from read-to-write tasks. As part of the field testing and validation of the prototype integrated task on the internet-based TOEFL, Cumming et al. (2005) compared the products of two independent (writing only) and four integrated (two reading-writing and two listening-writing) tasks produced by 36 test takers, focusing on discourse features such as lexical and syntactic complexity, grammatical accuracy, rhetoric, pragmatics, and verbatim uses of source text. The researchers found significant differences between the three task types in terms of lexical sophistication, syntactic complexity, argument structure, voice in source evidence, and message in source evidence. However, the two integrated tasks were similar to each other in many of the discourse analyses that the researchers conducted. Similarly, Gebril and Plakans (2009) analyzed 131 products of a reading-writing test for discourse features and source text use as well as investigated the writers' self-reported process through a post-test questionnaire with a particular focus on use of source texts. The researchers found statistically significant differences across proficiency levels in fluency, grammatical accuracy, indirect source use, and total source use, but no differences for lexical sophistication, syntactic complexity, direct source use, and direct source use without quotations. In the same research context, Gebril and Plakans (2013) holistically rated and analyzed 136

essays from a reading-based writing task and found fluency to be the only discourse feature that distinguished essays across all three levels, while grammatical accuracy and source use played a large role in distinguishing between the two lower levels. The authors suggest that broader textual features, including cohesion, content, and organization, may play an important role in the higher levels. They also argue that the integrated writing construct needs to include reading ability and discourse synthesis.

A more narrowly focused study that investigated the effect of specific linguistics features of essays on essay scores is Guo, Crossley, and McNamara (2013). This study identified seven features that most significantly predict scores from the TOEFL iBT integrated task: textual length, past participle verbs, word familiarity, verbs in 3rd person singular form, semantic similarity, verbs in base form, and word frequency. These seven predictors combined explained 53.3% of the variance in the scores.

Several other studies of test products focused more specifically on source text use. Ohkubo (2009) investigated the use of source texts in six test takers' written products for the integrated task in the TOEFL iBT through discourse analysis and post-test interviews. Weigle and Parker (2011) is another study that analyzed source text borrowing in essays from a reading-writing test. Most recently, Plakans and Gebril (2013) analyzed 480 essays from the integrated listening and reading to write task on the TOEFL iBT and found that (a) the importance of ideas from source texts, (b) use of the reading text, (c) use of the listening text, and (d) verbatim source use explained more than 50% of the variance in test scores. Specifically, higher scoring essays tended to use important ideas from the source texts and used the listening text as instructed, while the lower scoring essays tended to rely on the reading text and included more direct copying of source language.

One recent study by Sawaki, Quinlan, and Lee (2013) used discourse features of essays to investigate the factor structure of the TOEFL iBT integrated writing task. Confirmatory factor analysis was used to identify a model for the integrated task that has writing as the higher-order factor and three first-order factors, which are sentence conventions, productive vocabulary, and content. Additionally, the authors identified a comprehension higher-order factor model, which has comprehension as the higher-order factor and allows listening and reading skills to be included along with the three writing factors from the previous model.

A third vein of research on integrated writing tests is the perceptions of test users. Test takers' perceptions were investigated in pilot studies of integrated reading-writing tasks during the prototyping stage of the development of the internet-based TOEFL (Chapelle, Enright, & Jamieson, 2008). Kim (2008) collected evidence to validate a reading-to-write test as a diagnostic test in an academic writing course for international graduate students by investigating the perceptions and evaluations of students and instructors through questionnaires and interviews. Most recently, Weigle and Montee (2011) looked at rater perceptions of textual borrowing in integrated writing tests.

Other veins of research on integrated writing tests include generalizability studies and decision studies that use generalizability theory to examine the impact of number of tasks, number of raters, and rating designs on the reliability of writing scores. These studies stem from the concern over tasks and raters being the two main sources of score variability in performance assessment (Lee & Kantor, 2007, p. 354). Using listening-writing, reading-writing, and independent writing tasks, Lee and Kantor (2007) found that increasing the number of tasks is a more efficient way than increasing the number raters per essay to increase score reliability. Gebril (2009) compared reading-to-write tasks and independent tasks and found that the two task

types yield comparable score generalizability and similar score reliability. Furthermore, Gebril (2010) concluded from using the same data as the previous study that a combined score of the two task types is as reliable as scores obtained from either task. Two rating designs—having different raters score each task type and having the same raters score both task types—were found to produce scores with similar values of reliability.

Source-based writing tests are also used in the L1 context, although not necessarily for assessing writing skills per se. For example, Blattner and Fraziere (2002) use a writing test to assess critical thinking skills in which six short texts are provided as reading materials and test takers are required to use material from at least two readings in their essays. Furthermore, although from the field of physics education, Priemer and Ploog (2007) is probably the study that is conceptually the most similar to the current study in that the student participants had full access to the internet and used information from internet sources during a text production task on topics in physics. Since the text production task was viewed as a physics learning activity, the focus was on the content of the essays and students' learning of physics knowledge rather than writing skills. Nevertheless, the authors identified two groups of participants in terms of text production: "compilers" and "authors." The former group of students copied and pasted information from internet sources into their essays, while the latter group wrote all or most of their essays by themselves. In terms of methodology, an unidentified software program was used in the study which created log files of each participants' "every key stroke, written text, web page visited, time on a web page, etc" (p. 617).

Even though numerous studies have been conducted so far on the processes, products, test users' perceptions, and reliability of integrated reading-to-write tests, there have been few if any studies that looked at assessment of writing from internet sources in particular. The reading-

to-write construct, as it has been defined and researched in the previous writing assessment literature, is limited to the use of one or more printed texts that are pre-selected and directly presented to the test takers. This is insufficient in incorporating the use of internet sources, particularly test takers' independent searching and choosing of sources on the internet, which is an important and necessary skill for college student writers of the present age. The proposed web-search-permitted integrated writing test is my attempt as a test developer to define a new construct of academic writing, web-researching-to-write, that adds the dimension of internet literacy to the previously existing construct of reading-to-write by assessing the test takers' use of web sources in an integrated writing test. The test was also developed with the intention of using it as a final exam in an academic writing course for international undergraduate students to infer whether students have acquired the skills that were taught in class regarding use of web sources in writing as well as other general aspects of essay writing such as organization, supporting details, expression, and correctness. In the following, I present an interpretive argument for the use of scores from the web-search-permitted and web-source-based integrated writing test that will later be backed by evidence collected through the dissertation study and presented in the form of a validity argument.

An Interpretive Argument for the Web-Search-Permitted Integrated Writing Test

The web-search-permitted and web-source-based integrated writing test is a classroom-based test that I have developed to be used as a final exam in English 101C, which is the second of a two-course sequence of ESL academic writing courses for international undergraduate students at Iowa State University (ISU). New international students that do not meet the English language requirements are required to take a placement test before the beginning of their first

semester of studies. The placement test puts students into three levels. Students placed into the lowest and middle levels must take two and one ESL academic writing course respectively before proceeding to first-year composition courses (English 150 and 250). Students placed into the highest level are considered ready to take first-year composition with native-speakers of English and thus are not required to take any ESL courses. The two ESL writing courses play the role of orienting international undergraduate students to the academic writing conventions of English and equipping students with basic essay writing and grammar skills for writing. The syllabus for English 101C identifies the course goals and objectives as follows:

> The purpose of English 101C is to help ISU students increase their writing skills. The course will help students develop a mature writing style and an ability to integrate ideas, personal experiences, and external sources into their own writing. The course will further emphasize writing as a process and help students learn to improve writing through revision and editing workshops.
>
> *Upon completion of this course, students will be able to:*

- read challenging texts that reflect important themes and demand critical thinking;
- summarize and critique examples of mature writing styles and techniques;
- revise through multiple drafts to complete successful essays;
- construct coherent essays based on reading, interpreting, analyzing, critiquing, and synthesizing texts;
- adapt the structure, content, and tone of their writing to the knowledge and attitudes of their audience;
- use vivid, concrete language; concise, varied sentences; unified, cohesive paragraphs; gender exclusive English; and a college-level vocabulary; and

- proofread, edit, and correct their final copy for common errors of spelling, punctuation, capitalization, and usage.

At present, all of the English 101C sections must give a final exam during a two-hour period allotted by the university, and the test task is to write either a persuasive essay or a response essay depending on the instructor's decision. The possible prompts, including topics, length requirement, and evaluation criteria, are provided in Appendix A. Having taught English 101C for four semesters, however, I felt that the final exam should go beyond the writing of a simple persuasive essay or a reflection paper because students have already demonstrated their narrative writing skills in the first essay assignment, a personal essay, and their persuasive writing skills in the fourth essay assignment, an argumentative essay. Furthermore, my syllabus includes classroom activities and tasks that teach students how to (a) search for sources online through the library databases and search engines, (b) evaluate the credibility of online sources, and (c) use sources in writing by summarizing, paraphrasing, and quoting and by using in-text citations and references lists. There is also an in-class activity where students search for and compile a list of online writing resources such dictionaries, thesauruses, and grammar checkers. However, there is no formal essay writing assignment that requires source use, which led me to the idea that the final exam could be used as an opportunity for students to demonstrate all of the skills that they had acquired over the semester regarding source-based writing. Therefore, I developed and piloted a test task in Spring 2011 with a section of English 101C taught by another instructor and used the test again in Fall 2011 and Spring 2012 in my own three sections of 101C.

My test consists of one essay writing task that differs from the current English 101C final exam in terms of the prompt, topic, and evaluation criteria. The task was designed to maximize

authenticity by designing it to make the test performance match the target performance as much as possible. My task asks test takers to compose an argumentative essay using internet sources that they have searched for and selected during the test. Test takers are also allowed to consult online writing resources. The task tries to draw out test-taking behavior that resembles the processes that would normally be involved in source-based writing but within a fairly limited amount of time (two hours).

The construct that my test is intended to measure is "web-researching-to-write" or "source-based academic writing ability," with "source" referring specifically to "internet sources." My approach to construct definition is interactionist, as I am interested in both the trait (writing) and context (e.g., academic English [style], argumentative essay [genre], internet access to web sources [condition]) as well as the strategies required for their use.

It is worth developing an interpretive argument for this test because the test format is fairly novel. Since I am the first English 101C instructor to develop and implement such a test, and the test is going to be used to assign final exam grades in 101C, I would like to make sure that I am doing something that is beneficial for both the students and me as an instructor. I first developed and piloted one test task, and then followed a more rigorous and systematic test development and revision process to ensure that useful and meaningful evidence as backing for the interpretive argument could be accumulated along the way.

The data used to support the interpretive argument include pilot data that were collected in May 2011, test data collected in December 2011 and May 2012, and follow-up data collected in March-May 2013. The data include test-taking process data (Camtasia screen capture recordings), test product data (essays), test-taker perception data (post-test student questionnaires and interviews; follow-up student questionnaires and interviews), instructor perception data

(expert judgment interviews with English 101C instructors), and artifacts (syllabi, textbooks, and assignment sheets).

The primary intended audience for the interpretive argument is my POS committee, but other English 101C instructors and the ESL coordinators may also read the plan if they are interested in adopting the test for their own use. The committee members will be able to understand the process I went through to plan and develop the writing test, while other readers will be able to see the potential usefulness of the test for their teaching and assessment purposes.

An Overview of Test Interpretations, Uses, and Consequences

Scores from the integrated writing test have several intended meanings associated with them. First, according to Kane (2006), "a semantic interpretation draws conclusions based on assessment results" and "assigns meanings to these results" (p. 51). The primary semantic meaning that I am addressing in my interpretive argument is the following test score interpretation. The final essay scores are intended to reflect the extent to which students are able to write an argumentative essay by integrating information obtained from internet sources that they have searched for during the test and while making use of online language help options. The interpretations that are going to be made on the basis of the test scores pertain to how well students can demonstrate the source-based writing skills that they have learned and practiced in the English 101C classes.

Furthermore, Kane (2006) explains that a "decision procedure implements a policy and involves choices about what to do; it is evaluated in terms of its outcomes, or consequences" (p. 51). Therefore, the policy meanings that I am addressing in my interpretive argument are related to the decisions that will be made based on the test scores and the intended consequences of test

use. The test use that needs to be supported by my validity argument is the assignment of differential final exam grades as the test is used for the purpose of summative assessment at the end of a semester of ESL academic writing instruction. The final essay score will account for 10% of the final course grade. Therefore, the decisions to be made on the basis of test scores are relatively low-stakes because it is not very likely for a student to fail the class due to bad performance on the test.

There are several intended consequences of the use of my test. Firstly, students will get an idea of how much they have learned and improved after taking the course based on their performance on the test and after seeing the evaluation of results. The instructor, on the other hand, will get an idea of how much students have internalized the skills taught in the course. Furthermore, instructors and students will see the importance of internet literacy and source use in writing in the academic context. Another intended consequence is that test use would promote positive washback on teaching, so that instructors will pay more careful attention to teaching their students how to search for sources, evaluate sources, and incorporate information from sources into their essays.

Approach to Validation: An Argument-Based Approach

The argument-based approach to validation, which was initially conceptualized by Kane (1992, 2006, 2012, 2013) and applied to language testing by Chapelle, Jamieson, and Enright (2008) and Bachman and Palmer (2010), serves as the framework for my dissertation. Test validation has been defined by Messick (1989) in his seminal chapter as the collecting of evidence that supports the use of a test and the interpretations of the test scores from a variety of perspectives. Messick's ideas have been further developed by Kane (2006) through his proposal

of the argument-based approach to test validation in which claims about test score interpretations are supported by assumptions and backing. It is a pragmatic and praxis-oriented approach to validation (Chapelle, 2011) that "build[s] upon Messick's seminal work to provide praxis-oriented concepts and tools for planning, conducting, and interpreting validation research" (Chapelle, 2012b, p. 3). Chapelle, Enright, and Jamieson (2010) argue that the validity argument approach of Kane provides a clear framework for validation of language tests. This is demonstrated by Chapelle, Enright, and Jamieson (2008) through the construction of a validity argument for the new internet-based TOEFL, which made use of the interpretive argument-validity argument sequence.

The argument-based approach is an analytical approach that allows a researcher to propose and define the use of a test first and then conceptualize the construct that is accordingly being measured by the test. It is a use- or context-driven approach, which means that the focus is on the use of scores from a test or the context of test score use. The construct can be conceptualized by the researcher according to the intended use of test scores or the context of test score use. The approach "allows for validity arguments to be developed in a number of different ways and allows researchers to conceptualize the construct and the validation needs as they see fit" (Carol Chapelle, personal communication). The researcher makes the "argument" that the test is appropriate for the proposed test use and for measuring the conceptualized construct using a chain of inferences and evidential data. An interpretive argument allows researchers to outline an explicit chain of inferences connecting the domain and observation of test performance to the uses of the score assigned to the test performance and the decisions made based on the score. By evaluating the strength of the argument and the data and evidence that were collected to support each of the inferences that make up the argument, it is possible to identify weak links in the

interpretive argument. The argument-based approach makes explicit the things that may already be done implicitly in construct validation as defined by researchers in the construct-driven approach.

The core of the argument-based approach is the construction of an interpretive argument and a validity argument. These two arguments comprise the two-part process of validation. The interpretive argument is the product of the first phase of test validation (Briggs, 2004; Kane, 2006). It is an outline of the "inferences and assumptions that underlie score interpretation and use" (Chapelle, Enright, & Jamieson, 2008, p. 5). Thus, "it provides a framework for test development by indicating the assumptions that need to be met" (Kane, 2006, p. 60). The interpretive argument also serves the function of providing "a framework for the validity argument" because the "evidence called for in the validity argument is that needed to support the specific inferences and assumptions in the interpretive argument" (Kane, 2006, p. 60). Therefore, the interpretive argument helps specify the research needed to collect evidence for the validity argument. The interpretive argument can and should be constructed in the very early stages of or even prior to the test development process. In fact, Chapelle, Enright, and Jamieson (2008) suggest that "a draft interpretive argument would ideally be on the table during test design and field testing" (p. 23).

As the test is designed and becomes operationalized, the test developer can provide evidence from research, test data, and the test development process itself to support the claims in the interpretive argument. The validity argument is created by adding supporting evidence to the interpretive argument. It is "ultimately built through critical analysis of the plausibility of the theoretical rationales and empirical data that constitute support for the inferences of the interpretive argument" (Chapelle, Enright, & Jamieson, 2008, p. 5).

I am adopting an argument-based approach for the validation of my test score use and interpretations because of its benefits and strengths. First of all, the argument-based approach is useful as a framework that helps to organize the entire research process. By laying out the interpretive argument for the use of scores from a test, I can identify what types of evidence are needed as backing that would support the assumptions under the inferences in the argument. This in turn can help me envision how I can go about collecting the necessary evidence through research.

Chapelle (2008) contrasts the argument-based approach to validation with "the accumulation-of-evidence approach that derives from the past work in educational measurement (Cronbach & Meehl, 1955; Messick, 1989) and language assessment (Bachman, 1990; Weir, 2005)" (p. 321). She argues that the latter approach "can be problematic because of the difficulty in deciding what kind of evidence to gather and how much evidence is enough" (p. 321). However, within the argument-based approach, an "interpretive argument consisting of different types of inferences provides guidance as to the types of research needed" (Chapelle, 2008, p. 321). AERA, APA, & NCME's (1999) *Standards for educational and psychological testing* provide another approach to validation based on the five sources of validity evidence, but Sireci (2009) points out that the 1999 *Standards* provide a framework that is "not overly prescriptive" and that the "lack of strong, prescriptive rules for what needs to be done to "validate" an inference has frustrated some researchers and test evaluators because it is not clear when sufficient evidence has been gathered" (p. 31). Moreover, Chapelle et al. (2010), after comparing the *Standards* to Kane's argument-based approach, argued that the latter provides a clearer framework for validation by "offering specific guidance and conceptual infrastructure" (Chapelle, 2012a, p. 20).

The "network or chain of inferences…sets the agenda for validation" (Kane, 2012, p. 10), and this is what leads to the first major advantage of the argument-based approach. The interpretive argument clarifies and even dictates what kinds of data need to be collected as evidence in support of each inference. This can be very helpful for researchers in the planning stages of validation projects. Kane (2012) also suggests that the "interpretive argument is particularly useful in providing guidance in allocating research effort and in gauging progress in the validation effort" (p. 10). As a result, the researcher can focus his or her efforts on collecting the most relevant and problematic validity evidence to support the inferences and assumptions in the interpretive argument.

Secondly, the argument-based approach is useful as a rhetorical device for the structure of the dissertation. A major advantage of an argument-based approach to validation is the coherence it provides to the readers. Readers are taken inference by inference from the target language use domain to a sample performance to an observed score to an expected score to a construct to a target score to score use and finally to consequences of score use. At each step, the readers can make a judgment of whether the evidence fully supports the inference being made. The research questions can be identified from the types of evidence required as backing for assumptions associated with each inference. The results section can be structured around the inferences in the interpretive argument and the pieces of evidence that support the assumptions. Chapelle (2008) posits that the validity argument can be both "an integrated structure to organize test-related research results that pertain to validity and a clearly articulated argument to translate into terms that a broader range of constituents can understand" (p. 350). In a validity argument, the "backing is expressed through statements that summarize findings that support inferences" and these statements compose "an overall argument leading to the intended conclusion" (p. 321).

In short, an argument-based approach to validity provides a framework that can combine, organize, and present all of the bits and pieces in one coherent "story" (Chapelle et al., 2008, p. 23).

Thirdly, the argument-based approach makes it easy for me to identify at the end of my study what additional backing is needed to strengthen the validity argument. The gaps that appear in the validity argument directly point to further research or data needed to strengthen the backing in support of certain assumptions in the argument.

Especially for test developers wishing to show that the use and interpretation of scores from a test are valid, a more comprehensive framework like the interpretive argument and validity argument can be an effective way to inform readers how the test was conceptualized, developed, and implemented in a specific context of language use. There could certainly be weaknesses in the validity argument due to the lack of available data or conflicts with the realities of the test use context, but having those weaknesses be clearly visible could actually make it easier for the test developers to outline further research that needs to be conducted to improve the robustness of the validity argument.

My roles in the test validation project can affect the type of validity argument that would ultimately result from the work with my interpretive argument. As the test developer, my aim is to advocate and support the interpretations and uses of the test, but at the same time, I am also a test user who evaluates the interpretations. Thus, I will be taking both a confirmationist stance and an evaluative stance to validation. My main tasks in constructing the interpretive argument would be to develop the inferences, warrants, and assumptions and to suggest backing to support the assumptions. However, I will also strive to take a critical and evaluative stance of inquiry in

order to identify and suggest potential rebuttals to the inferences in my interpretive argument whenever possible.

The Interpretive Argument

In this section, the claims, inferences, warrants, and assumptions that make up the interpretive argument will be outlined and presented. The backing that is needed as evidence to support the assumptions will also be described, along with potential rebuttals. I used Chapelle, Enright, and Jamieson (2008) as the major source of ideas for the content and presentation of my interpretive argument, while a number of claims and warrants have been adapted from Bachman and Palmer's (2010) assessment use argument framework.

In Figure 1, a specific imaginary student's test performance is used to illustrate the grounds and claims of the interpretive argument, along with the inferences that link each of the grounds to its respective claim. There are a total of seven inferences and eight grounds/claims. The chain of inferences connects the target language use domain and test scores to the intended test use and consequences. Each inference connects grounds or data to a claim or an intermediate conclusion. The claim becomes the data for the following inference.

A summary of the inferences, warrants, assumptions, and backing in the interpretive argument is provided in Table 2.1 An explanation of the interpretive argument will follow.

Consequences:
• Test use will have positive washback effects on how English 101C is taught.
• The student will have a better understanding of the importance of internet literacy and appropriate source use in writing.

↑ Implication

Test use: The student receives a high enough final exam grade to pass English 101C.

↑ Utilization

Target score: The student is likely to obtain high scores on other indicators of source-based academic writing ability, including self-assessment and instructor judgments about performance in courses that require source-based writing.

↑ Extrapolation

Construct: The student has ability to apply the knowledge and skills learned in English 101C regarding internet literacy, source use, and essay writing.

↑ Explanation

Expected score: The student is likely to receive a score of "A" on other source-based academic writing measures or assignments.

↑ Generalization

Observed score: The student's essay received a score of "A."

↑ Evaluation

Observation: A student wrote within two hours a well-organized argumentative essay that incorporates one or more quotes, paraphrases, and/or summaries from credible internet sources.

↑ Domain description

Target domain: Writing an argumentative essay using web sources requires internet literacy and source-based academic writing ability. The skills that constitute these abilities are taught through numerous tasks and in-class activities in English 101C.

Figure 1. An illustration of the grounds, claims, and inferences in the interpretive argument.

Table 2.1

Summary of the Inferences, Warrants, Assumptions, and Backing in the Interpretive Argument

Inference in the Interpretive Argument	Warrant Supporting the Inference	Assumptions Underlying Warrant	Backing Sought to Support Assumption
Domain description (Target language use domain → observation of performance (essay))	Observations of performance on the integrated writing test reveal relevant skills, knowledge, abilities, and processes in situations representative of those in the target domain of web-source-based academic writing in college courses, particularly the knowledge and skills taught in English	Critical English language skills, knowledge, abilities, and processes needed for source-based academic writing in English-medium university classes can be identified.	Domain analysis (expert consensus, syllabus and textbook analysis)
		Possible assessment tasks that are representative of the domain can be	Domain analysis (expert consensus, assignment sheet analysis)

	101C.	identified.	
		An assessment task that requires important skills and is representative of the domain can be simulated.	Systematic process of task design and modeling
Evaluation (Observation of performance → observed score)	Observations of performance on the integrated writing test are evaluated to provide observed scores reflective of targeted language abilities (web-source-based academic writing).	Task administration conditions are appropriate for providing evidence of targeted language abilities.	Multiple task administration conditions were developed, trialed, and revised.
		The rubric for scoring essays is appropriate for providing evidence of source-based academic writing ability and has been applied as intended.	Systematic rubric development
		Instructors can be trained to avoid bias for or against different groups of students.	Rater training and calibration
Generalization (Observed score → expected/universe score)	Observed scores are estimates of expected scores over the relevant parallel versions of tasks and within and across raters.	Task and rating specifications are well defined so that parallel tasks can be created.	Systematic development of test spec for producing parallel tasks
		Different ratings by the same instructor are consistent.	Intra-rater reliability
		Ratings of different instructors are consistent.	Inter-rater reliability
Explanation (Expected/universe score → construct)	Expected scores are attributed to a construct of web-source-based academic writing ability, which is defined by the English 101C syllabus and the teaching/learning activities in the class. The interpretations about the students' ability to search for and select internet sources and incorporate information from the sources in an argumentative essay are meaningful with respect to the English 101C teaching syllabus and the teaching/learning activities in the class.	The characteristics of the integrated writing test correspond closely to those of instructional tasks.	Comparative analysis of test task and instructional tasks in English 101C
		The criteria and procedures for evaluating the responses to the integrated writing test correspond closely to those that instructors have identified as important for assessing performance in other writing tasks in the instructional setting.	Comparative analysis of test rubric and rubrics used in English 101C
		The linguistic knowledge, processes, and strategies required to successfully complete the test are in keeping with theoretical expectations.	Examination of task completion processes (using Camtasia screen capture recordings) and discourse analysis (essays) supported the development

			of and justification for the task
		Test performance varies according to amount and quality of experience in learning (to write in) English.	Comparison studies of group differences (Examination of relationships between test performance and English learning, writing experience, and internet use)
Extrapolation (Construct → target score)	The construct of web-source-based academic writing as assessed by the integrated writing test accounts for the quality of web-source-based academic writing performance in college courses, particularly with regard to those skills taught in English 101C.	Performance on the test is related to other criteria of source-based writing ability in the academic context including self-assessment and future performance.	Criterion-related evidence (Examination of relationships between test performance and students' self-assessment of their own source-based academic writing ability and students' performance in a post-English 101C course that requires the completion of source-based writing assignments)
Utilization (Target score → test use/decision)	Estimates of the ability to search for and select internet sources and incorporate information from the sources in an argumentative essay, which are obtained from the integrated writing test, are useful for making decisions about final exam grades and appropriate curricula for students in English 101C. The summative decisions that are made about students' progress reflect relevant existing educational and societal values and relevant university regulations and are equitable for the 101C students. These decisions will be made by the 101C instructors.	Students have equal opportunities to learn or acquire the ability to write from internet sources in English 101C. The test scores provide useful and meaningful information to the students and instructors regarding students' source-based writing abilities, and the meaning of test scores is clearly interpretable by test takers and instructors.	Examination of perspectives of students Score descriptors are provided to students along with their test score; students' and instructors' perspectives on the usefulness, clarity, and interpretability of the descriptors
Implication (Test use → consequences)	The consequences of using the integrated writing test and the decisions that are made are beneficial to the students (test takers), the English 101C teachers, English 150 teachers who will teach these students in	The test will have a positive influence on how academic writing is learned and taught, that is, test use promotes positive washback effects on English 101C.	Washback studies (Instructor and expert interviews, follow-up student questionnaires and interviews)

	the following semester, and instructors of academic courses at ISU who will encounter these students in their classes. Stakeholders: 1) Students in the English 101C classes 2) The English 101C instructors 3) Other instructors who will teach the students in future semesters	Score reports are distributed to students in a timely manner.	Rater training and rating sessions

The first inference in the interpretive argument is domain description that links the target language use domain to the observation of performance, specifically the essay. This inference is supported by the warrant that observations of performance on the integrated writing test reveal relevant skills, knowledge, abilities, and processes in situations representative of those in the target domain of web-source-based academic writing in college courses, particularly those knowledge and skills that are outlined in the English 101C syllabus under course goals and taught in the course. There are three assumptions underlying this warrant: (a) critical English language skills, knowledge, abilities, and processes needed for source-based academic writing in English-medium college classes can be identified; (b) possible assessment tasks that are representative of the domain can be identified; and (c) an assessment task that requires important skills and is representative of the domain can be simulated. The first two assumptions will be supported by backing collected through domain analysis. The methods of expert consensus and document analysis of syllabi, textbooks, and assignment sheets will help identify the critical skills, knowledge, abilities, and processes needed in the target domain and an assessment task that can represent them. The potential experts would be instructors of English 101C, 150, and 250, the coordinators of both the ESL and first-year composition programs, and perhaps instructors of content courses at ISU who require source-based writing from students. The third

assumption will be supported by backing based on the systematic process of task design and modeling. I will present the test task I have developed to the experts and ask for comments and feedback on the appropriateness of the task for the purpose of the test.

The second inference is evaluation, which links the observation of performance to an observed score. This inference is supported by the warrant that observations of performance on the integrated writing test are evaluated to provide observed scores reflective of targeted language abilities. There are three assumptions underlying this warrant: (a) task administration conditions are appropriate for providing evidence of targeted language abilities; (b) the rubric for scoring essays is appropriate for providing evidence of source-based academic writing ability and has been applied as intended; and (c) instructors can be trained to avoid bias for or against different groups of students. The backing for the first assumption is that multiple task administration conditions were developed, trialed, and revised. Different conditions that were trialed include wording of the prompt, delivery mode of the prompt (Moodle quiz or email), and time limit (90 minutes or 120 minutes). Backing for the second assumption will come from systematic rubric development. The rubric will be developed and revised based on expert consensus of important, relevant criteria. Lastly, backing for the third assumption will come from the provision of rater training and calibration.

The third inference in the interpretive argument is generalization, which links the observed score to an expected score. This inference is supported by the warrant that observed scores are estimates of expected scores over the relevant parallel versions of tasks and within and across raters. There are three assumptions underlying this warrant: (a) task and rating specifications are well defined so that parallel tasks can be created; (b); different ratings by the same instructor are consistent and (c) ratings of different instructors are consistent. The first

assumption will be supported by the backing that a test specification was systematically developed for the production of parallel tasks. The second and third assumptions will be supported by examination of intra-rater reliability and inter-rater reliability, respectively. One instructor would rate a set of essays once and then rate the essays again after a period of time to calculate intra-rater reliability, while two or more instructors would rate the same set of essays to calculate inter-rater reliability.

My interpretive argument might be the most vulnerable to rebuttals against the generalization inference because in the operational use of the test, there is only one task and one rater, who is the instructor of the class. A potential rebuttal would be that one task is not large enough of a sample of performance and would thus lower the generalizability of scores. This rebuttal comes, in part, from Kane's (2006) observation that "generalizability over performance tasks cannot be taken for granted" (p. 57). Rebuttals can also come from empirical research findings which indicate that "essay graders generally find it difficult to maintain consistent standards over time" (Kane, 2006, p. 50). In defense against these rebuttals, I could make the argument that if the test is implemented in the future across multiple sections of English 101C, several parallel tasks can be used in the same semester, and there could be additional raters to help increase inter-rater reliability. Furthermore, findings from my study may be able to show that (a) raters are consistent, both within and across raters, and (b) one test task is sufficiently representative of the domain.

The fourth inference is explanation, which links the expected score to the construct. It is supported by the warrant that the interpretations about the students' ability to search for and select internet sources and incorporate information from the sources in an argumentative essay are meaningful with respect to the English 101C syllabus and the teaching/learning activities in

the class. In other words, expected scores are attributed to a construct of web-source-based academic writing ability. There are four assumptions underlying this warrant: (a) the characteristics of the integrated writing test correspond closely to those of instructional tasks; (b) the criteria and procedures for evaluating the responses to the integrated writing test correspond closely to those that instructors have identified as important for assessing performance in other writing tasks in the instructional setting; (c) the linguistic knowledge, processes, and strategies required to successfully complete the test are in keeping with theoretical expectations; and (d) test performance varies according to amount and quality of experience in learning (to write in) English. Backing for the first assumption will come from comparative analysis of the test task and instructional tasks, while for the second assumption, backing will come from comparative analysis of the test rubric and rubrics used for essay assignments in English 101C. The third assumption will be supported by the backing that examination of task completion processes using Camtasia screen recordings and post-test interviews and discourse analysis of the essays support the development of and justification for the test task. Lastly, the fourth assumption will be supported by comparison studies of group differences through the examination of relationships of test scores with English learning, writing experience, and internet use investigated through student questionnaires.

Extrapolation, which is the fifth inference in the interpretive argument, links the construct to the target score. The warrant that supports this inference is that the construct of web-source-based academic writing as assessed by the integrated writing test accounts for the quality of web-source-based academic writing performance in college courses, particularly with regard to those skills that are identified in the English 101C syllabus and taught in the class. The assumption underlying this warrant is that performance on the test is related to other criteria of source-based

writing ability in the academic context. Backing will be sought from criterion-related evidence which examines the relationships between test performance and (a) students' self-assessment of their own source-based academic writing ability and (b) students' performance in a post-English 101C composition course which requires the completion of source-based writing assignments.

The sixth inference in my interpretive argument is utilization, which links the target score to test use. This inference is supported by the warrant that estimates of the ability to search for and select internet sources and incorporate information from the sources in an argumentative essay, which are obtained from the integrated writing test, are useful for making decisions about final exam grades and appropriate curricula for students in English 101C. Furthermore, the summative decisions that are made about students' progress reflect relevant existing educational and societal values and relevant university regulations and are equitable for the 101C students. These decisions will be made by the 101C instructors. Two assumptions underlie the warrant: (a) students have equal opportunities to learn or acquire the ability to write from internet sources in English 101C, and (b) the test scores provide useful and meaningful information to the students and instructors regarding students' source-based writing abilities, and the meaning of test scores is clearly interpretable by test takers and instructors. Backing for the first assumption will come from perspectives of students. The second assumption will be supported by the backing that score descriptors are provided to students along with their test scores. Further backing will come from students' and instructors' perspectives on the usefulness, clarity, and interpretability of the descriptors.

The seventh and final inference is implication, which links the test use to the intended consequences. The warrant that supports this inference is that the consequences of using the integrated writing test and the decisions that are made are beneficial to the students (test takers),

the English 101C teachers, English 150 teachers who will teach these students in the following semester, and instructors of academic courses at ISU who will encounter these students in their classes. The stakeholders, therefore, are (a) students in the English 101C classes, (b) the English 101C instructors, and (c) other instructors who will teach the students in future semesters. The two assumptions that underlie the warrant are that (a) the test will have a positive influence on how academic writing is learned and taught, that is, test use promotes positive washback effects on English 101C; and (b) score reports are distributed to students in a timely manner. The backing for the first assumption will be gathered from washback studies using instructor interviews, student questionnaires, and student interviews, while the backing for the second assumption will come from the rating sessions by measuring and recording the time it takes to rate each essay.

Conclusion

The overall approach to gathering backing in support of the interpretive argument is to begin with the first inference (domain description) and go through the entire interpretive argument inference by inference. Since I began with the uses, interpretations, and intended consequences and worked my way down the ladder in the test development process as suggested by Bachman and Palmer (2010), I will now begin the validation process with the domain description and observations of test takers performance and work my way back up the ladder to build the validity argument. This is also in accordance with Kane's (2006) analogy of crossing the inference bridges one by one with a warrant or ticket that is supported by backing.

Research Questions

The first purpose of the dissertation study is to investigate how the newly defined construct of web-researching-to-write affects the processes, products, and perceptions of test users. Specifically, the study aims to determine how students display internet literacy during the web-search-permitted integrated writing test, how students incorporate internet sources into their essays, and how students and other stakeholders perceive the additional requirement of using internet sources in a writing test. The second purpose of the study is to collect evidence that supports the use of scores from the web-search-permitted integrated writing test as a final exam in English 101C, an academic writing course for international undergraduate students at Iowa State University. The findings from the study will become backing to support the inferences that are proposed in the interpretive argument, and as a result, the backing and interpretive argument will be incorporated into a validity argument.

To fulfill the above two purposes, based on the seven inferences and 18 types of required backing identified in the interpretive argument, the following seven sets of research questions were formulated:

1. Domain description inference
    1.1. Domain analysis (skills, knowledge, abilities, and processes) – What are the important skills, knowledge, abilities, and processes needed for source-based academic writing in college courses as identified by experts, syllabi, and textbooks?
    1.2. Domain analysis (possible assessment tasks) – What are possible assessment tasks that are representative of the domain of source-based academic writing in college courses as identified by experts, syllabi, and textbooks?

1.3. Systematic process of task design and modeling – How much did experts think that the web-search-permitted integrated writing test samples important skills and is representative of the domain?

2. Evaluation inference
   2.1. Multiple task administration conditions – How did the test takers feel about the test administration conditions (instructions and time limit)?
   2.2. Systematic rubric development – What did experts think about the appropriateness of the rating rubric for providing evidence of web-source-based academic writing ability?
   2.3. Rater training and calibration – How much can instructors be trained to avoid bias for or against different groups of students?
3. Generalization inference
   3.1. Systematic development of test specification for producing parallel tasks – How much did experts find the test task specification well defined for producing parallel tasks?
   3.2. Intra-rater reliability – How high is the intra-rater reliability?
   3.3. Inter-rater reliability – How high is the inter-rater reliability?
4. Explanation inference
   4.1. Comparative analysis of test task and instructional tasks – How much does the test task reflect instructional tasks in English 101C?
   4.2. Comparative analysis of test rubric and course rubrics – How much does the test rubric reflect the rubrics used to evaluate writing in English 101C?

4.3. Test completion processes and discourse analysis of products – What test-taking processes did test takers follow, and what web-searching behaviors did test takers show? What online language help tools did test takers consult? What relationships are there between test-taking processes and test scores? Do the test scores reflect how well web sources are used in the essays? How do the selection of sources, attribution to sources, and integration of source language relate to scores or differ across score levels?

4.4. Comparison studies of group differences – How is test performance related to test takers' English learning, writing experience, and internet use?

5. Extrapolation inference

5.1. Criterion-related evidence – How is test performance related to students' self-assessment of their source-based academic writing ability and students' performance in a post-English 101C course which requires the completion of source-based writing assignments?

6. Utilization inference

6.1. Equal opportunity to learn – How equal did test takers perceive the instruction and preparation they received before the test?

6.2. Usefulness, clarity, and interpretability of score descriptors – What did test takers and experts think about the usefulness, clarity, and interpretability of the score descriptors?

7. Implication inference

7.1. Washback studies – What are the washback effects of test use on instruction and learning?

7.2. Controlled rating time and timely distribution of score reports – How long does it take for raters to rate an essay? How long does it take for raters to rate essays for two course sections?

CHAPTER 3

METHODOLOGY

Research Design

This study used a mixed methods research design in which both qualitative and quantitative data are collected and analyzed. Creswell and Plano Clark (2007) identify four types of mixed methods research designs: triangulation, embedded, explanatory, and exploratory. The current study adopted the triangulation design. This design was chosen because the research questions were answered by results from both quantitative and qualitative data analyses. Some questions were answered based on the analysis of both qualitative and quantitative data. Others were answered based on only qualitative data or only quantitative data. When only qualitative data were available for a research question, attempts were made to quantify the data if possible.

Context

The context in which the web-search-permitted integrated writing test was used is English 101C, an academic writing course for international undergraduate students at Iowa State University. Students are placed into the course on the basis of scores that they obtain in an English placement test which is given to them on arrival to campus before the beginning of their first semester of studies. Specifically, the English placement test places students into three levels. Students placed into the lowest level must take two ESL writing courses (English 101B and 101C) before proceeding to mainstream first-year composition courses (English 150 and 250), whereas students placed into the middle level must take one (English 101C). Students placed into the highest level are considered ready to take first-year composition with native-speakers of English and are not required to take any ESL courses. English 101C is the second in the two-

course sequence of ESL writing courses for undergraduates. This means that students enrolled in this course have either (a) received the middle level grade in the placement test or (b) completed the first ESL writing course and are now in the second one.

The two ESL writing courses play the role of orienting international students to the academic writing conventions of English and equipping students with basic essay writing and grammar skills. The first course has a stronger focus on grammar in writing. The second course, which is of interest in the current study, places a stronger focus on general essay writing skills. Specifically during the two semesters in which the test was implemented, the course adopted a process-based writing approach that encouraged students to go through an iterative process of brainstorming (invention), drafting, peer response, revising, and editing when working on essays. Students were given four essay assignments based on essay type (personal, cause and effect, comparison and contrast, and argumentative) as well as journal assignments, vocabulary project assignments, and a final essay exam. The syllabus included units and activities on paraphrasing, taking notes on a text, summarizing, punctuating direct speech, plagiarism, library and internet research, quoting, and citing and documenting sources. Furthermore, the instructor of the sections that participants were enrolled in provided students with a list of online resources like dictionaries and grammar guides in the middle of the semester. However, none of the four essay assignments required the students to use outside sources.

Participants

The first group of participants is the test takers. Test takers were recruited from three sections of English 101C, all three of which were taught by the same instructor who was also the researcher. Two sections were taught in Fall 2011, while the remaining one section was taught in

Spring 2012. Out of a total of 71 students that were enrolled in the three sections and that took the final test as part of the English 101C curriculum, 50 students agreed to the collection of test data and participation in the study by signing the informed consent document before beginning the test. The 50 students consisted of 15 students from the first section, 16 students from the second section, and 19 students from the third section. The background information of the students is summarized in Table 3.1.

Table 3.1

Background Information of Test takers (N=50)

Category		**Number**
First language	Chinese	43
	Korean	2
	Spanish	2
	Arabic	1
	Congo	1
	Thai	1
Gender	Male	33
	Female	17
Major	Business/accounting/finance	20
	Engineering	15
	Math/statistics/sciences	11
	Design/architecture	2
	History	1
	Undecided	1

The majority of students spoke Chinese as their first language, while almost all students were majoring in business, engineering, math, or science. The students' English proficiency level was intermediate to high-intermediate as can be inferred from their self-reported proficiency test scores. The average scores were around 550 on the paper-based TOEFL, 79 on the internet-based TOEFL, and 6.3 on the IELTS. When the IELTS scores and PBT TOEFL scores were converted into iBT TOEFL scores using ETS's comparison chart and the MELAB/TOEFL Concordance

Table[1] respectively, the iBT TOEFL scores ranged from 69 to 97.5. The students also reported having learned English for almost 8 years on average and of having lived in the United States for a little less than one year on average. Among the 50 students, 6 participated in a post-test interview within four days of the test, and 9 participated in a follow-up interview two or three semesters after the test. Two students participated in both the post-test and follow-up interviews.

The second group of participants consists of domain experts and potential stakeholders: five previous and current instructors of English 101C and 150. Four instructors had taught both English 101C and 150, while one had taught only English 101C. These participants provided expert judgment on numerous aspects of the test during individual interviews with the researcher. All five instructors were international graduate students in applied linguistics and technology, each with numerous years of ESL teaching experience and near-native English proficiency. Four instructors had additional one or two semesters of experience in teaching English 150.

In addition to the test takers and experts were six essay raters. Three raters came from the domain expert and stakeholder group described above. The other three were also international graduate students in applied linguistics and technology with extensive ESL/EFL teaching experience and near-native proficiency in English. These latter three raters were teaching assistants as well, but only one of the three had experience teaching English 101C and 150. The remaining two had taught and were teaching other writing courses for international students.

Materials and Instruments

First of all, a test task specification was written following Davidson and Lynch's (2002) framework (Appendix B). The specification has five components: general description, prompt

---

1 http://www.cambridgemichigan.org/sites/default/files/resources/MELAB_ConcordanceTable.pdf

attributes, response attributes, sample item, and specification supplement. The general description section includes a description of what is to be tested as well as the general and specific objectives of the test. The prompt attributes section describes what will be given to the test taker to prompt him or her to produce a response. In the case of the integrated writing test, this includes directions, the form of the task, and requirements for the selection of a topic. The response attributes section describes what response test takers will produce and how the response will be evaluated.

Based on the above test task specification, a prompt was developed. The prompt that was used for the test can be found in Appendix C. An argumentative essay topic was chosen because the fourth and last assignment for the writing course was an argumentative essay, and it was also assumed that students would be familiar with this type of writing from taking a standardized English proficiency test for admission to the university, an English placement test upon arrival at the university, and a diagnostic test at the beginning of the semester, all of which required the writing of a persuasive opinion-based essay. The prompt asks test takers to use information from internet sources to support their opinions and also informs them of the fact that they are allowed to use online help options.

The test was created as an open-ended response item in a Moodle quiz. Moodle is the course management system used by the writing courses at the university. Since the participants already had accounts in the system and were enrolled in the course Moodle, the Moodle quiz was imported into the existing course site as the final exam. Figure 2 shows a screenshot of the prompt as viewed by the students. The time remaining is always displayed near the top left corner of the screen as a floating clock. The prompt is followed by a text box below, in which students type their essays. Once students submit their essays by clicking on the submit button

below the text box, the instructor/researcher can view all of the submitted essays under the "Results" tab.

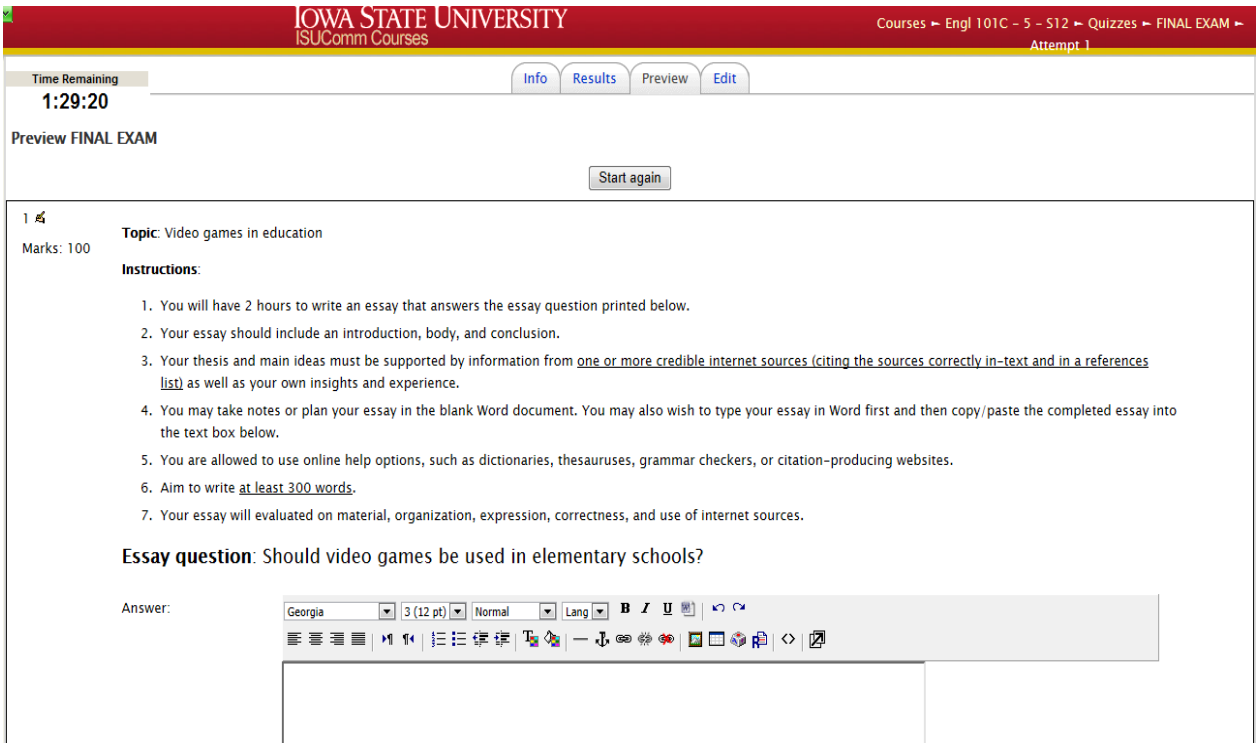

Figure 2. Screenshot of test prompt.

Other data collection materials include questionnaires and interview protocols. A post-test questionnaire was developed to record the test takers' perceptions of the test as well as collect information about the test takers' personal profile and background, standardized test scores, experience or patterns of internet use, and previous English writing experiences (Appendix D). The last item on the questionnaire asked test takers to choose a time and date if they were interested in being interviewed by the researcher about their test-taking experience. The post-test questionnaire was created using the questionnaire tool in the course Moodle. An interview protocol for post-test semi-structured interviews with students was also developed (Appendix E). The protocol included questions that ask about the test takers' experiences during

the test and perceptions of the test. A few of the questions were adapted from Gebril and Plakans (2009).

The follow-up questionnaire for test takers can be found in Appendix F. The items on this questionnaire were also used as the basis for follow-up interviews with test takers (see Appendix G for interview protocol). The follow-up questionnaire and interview protocol consisted of questions that asked students to recall their test experience at the end of English 101C and share their experiences in any post-101C composition courses such as English 150 and 250. The questionnaire and interview also asked students to read a copy of the essay rating rubric and indicate their perceptions of the score descriptors.

Separate interview protocols for individual interviews with experts and stakeholders were prepared (Appendix H). Questions on the protocol for interviews with previous and current instructors of English 101C were added with the purpose of obtaining expert judgment of the domain, sampling, test specification, test prompt, rating rubric, and score descriptors as well as opinions about source-based writing, use of the proposed test, and potential washback effects of the test on teaching and test preparation. Questions on the protocol for interviews with previous and current instructors of English 150 collected expert judgment of the domain, appropriateness of sampling, and usefulness of the test in preparing international undergraduate students for the demands and requirements of source-based writing in English 150.

Lastly, an essay rating rubric was adapted from two existing rubrics: (a) the rubric for the argumentative essay assignment in English 101C and (b) a rubric for an integrated writing test used within an English placement test at a U.S. university. In the pilot study, a holistic rating rubric (Appendix I) that was an adaptation of the latter of the two existing rubrics was used to divide the essays into four groups. However, to provide the test takers with more detailed

feedback on specific areas of writing and to more closely resemble the assignment rating rubrics in English 101C, I decided to use an analytic rating rubric in the current study (Appendix J and K). The final rubric is composed of four rows and five columns. In the leftmost column are four criteria which correspond to four major aspects of essay writing: material, organization, expression, and correctness. A brief description of each criterion is given in each of the four boxes of the leftmost column. In the remaining four columns are descriptors for each of four levels or grades (A, B, C, and D) for each of the four criteria. The four levels are assigned a different numerical score, with the highest scores for the four criteria adding up to 100. Use of web sources and citation of sources were incorporated into material and correctness respectively. The four criteria and their weights were retained to provide consistency with the other four major assignment grading rubrics used in English 101C which also used the same four criteria. It was also deemed that use of web sources is closely related to material and the development of content, while the citation of sources is a mechanics issue that belongs to the correctness criterion.

Procedure

Test Administration, Post-Test Questionnaire, and Post-Test Interviews

The test was administered as a final exam for the writing course in each of the classes. On the day of the test, test takers came to the computer lab where the test was to be given. They were seated at iMac computers that had been set up with a screen capturing program (Camtasia) running in the background. A web browser (Firefox) and a blank Microsoft Word document were displayed on the monitor. The web browser was pointing to the log-in page to Moodle. A cover sheet listing the steps for the participants to follow and an informed consent form were placed in front of each computer.

When all the participants were seated, they first read and signed the informed consent form. They were then asked to log into the course Moodle and proceed to the test whenever they were ready. The participants were given 120 minutes to read the prompt and construct an essay response. Immediately after the test, test takers were asked to complete the post-test questionnaire in Moodle to share their reactions to the test and experiences during the test as well as their previous experiences with computers and writing and their personal information. On the post-test questionnaire, students were also asked to sign-up for and participate in a voluntary interview to be conducted at a later time during the week following the test. The test session was two hours, but almost all students left before the two hours were over. Once all test takers had left the computer lab, the researcher saved the Camtasia screen recordings in mp4 format on memory cards.

The test administration procedure resulted in being somewhat different for one of the two sections in Fall 2011. The Moodle server was down on the morning of the final exam for one of the sections, and the researcher had to send the prompt in an email to the class. The students drafted their essays in a Word document or in their email text box and sent the completed essays either as an email attachment or email text. The test takers were given the same 120 minutes, and they had full access to the internet except the course Moodle. The Moodle server had been repaired toward the end of the test period, but since most of the test takers did not realize this, they left the computer lab after completing only the essay test. A few students who stayed long enough in the computer lab were able to complete the questionnaire within the test period. The researcher emailed the section right after the exam, requesting the students to respond to the post-test questionnaire in Moodle.

Six test takers who had expressed an interest in being interviewed on the post-test questionnaire were contacted within one day of the test to arrange a time. The interviewees did not know their test scores before coming to the interviews. The post-test interviews, which were held in the researcher's office or an empty classroom, lasted for 27 minutes on average, and ranged from 12 to 61 minutes in length. All interviews were recorded with a voice recorder to be later transcribed.

Follow-Up Student Questionnaire and Interviews

In the spring semester of 2013, all 50 previous 101C students who had participated in the study as test takers were contacted by email and were invited to participate in a follow-up interview with the researcher. Nine students responded and participated. Each interviewee came to an empty computer lab and was first asked to read and sign an informed consent form and to complete a paper-and-pencil questionnaire. The questionnaire took approximately 10 minutes to complete. The researcher then asked students for permission to record the interview, followed by semi-structured interview questions based on the student's questionnaire responses. The lengths of the interviews were 34 minutes on average and ranged between 21 and 51 minutes. Interviews were audio-recorded for later transcription.

Expert Judgment Interviews

Expert interviewees were contacted by email or in person to request their participation in an individual interview with the researcher. The interviews were held in an office or an empty computer lab and were audio-recorded for later transcription. Each interview lasted approximately one hour.

Data Analysis

The data consisted of (a) test-taking process data in the form of 48 Camtasia screen recordings, (b) product data in the form of 50 essays, (c) test-taker perception data in the form of 40 post-test questionnaire responses and 6 post-test interview recordings as well as 9 follow-up questionnaire responses and 9 follow-up interview recordings, (d) expert and stakeholder perception data in the form of 5 instructor interview recordings, and (e) artifacts in the form of syllabi, textbooks, and assignment sheets for English 101C and 150. Two Camtasia screen recordings were lost because two students logged out of the computer after finishing the test, despite the instructions on the cover sheet and the researcher's oral request. Ten post-test questionnaire responses were not collected, even after an email reminder requesting test takers to complete the questionnaire in Moodle was sent to non-responders immediately following the test.

Coding of Screen Capture Data

The screen capture data collected were analyzed to tap into the test-taking processes and strategies (RQ 4.3). The 48 screen capture files in mp4 format were viewed and coded by the researcher for search behaviors, key words searched, websites consulted, writing behaviors, and stages of the writing process. The time point within the video file was also noted for every action in an hour:minute:second format (e.g., 01:12:13). At the beginning of the coding process, a coding scheme that was developed in a pilot study (Appendix L) was used to code a few screen recordings. By the end of this preliminary coding stage, a final coding scheme (Appendix M) was developed, which was then used to code the rest of the recordings. The coding was done in a Microsoft Excel spreadsheet. There was no second coder for the analysis of screen recordings

because it was deemed that the coding did not involve enough subjective judgment to warrant the need for a second coder.

Rating and Discourse Analysis of Essays

The 50 essays were first copied and pasted from Moodle into a Microsoft Word document, and each essay was given a number from #1 to #50. The essays were then randomly rearranged within the Word document. Next, the essays were divided into five Word documents of 10 essays each, and each Word document was saved as a PDF file. After this preparation, the researcher sent each of the six second raters a PDF file containing 10 essays and an Excel rating sheet. Since there were a total of 50 essays and six raters, two raters were given the same set of 10 essays. Each rater was also sent a rating guide and rubric in PDF format (Appendix N) to provide rater training and calibration with four essays from a pilot study (RQ 2.3). The raters had to self-train by reading the rubric and the four benchmark essays.

The second raters each separately rated his or her set of 10 essays once according to the rubric. The researcher rated all 50 essays twice with six weeks in between to obtain two sets of ratings. The essays were rated by the researcher in a scrambled order the second time to further lessen the effect of memory and to ensure that the second ratings are independent of the first (Bachman, 2004, p. 169). In each set of ratings, four analytic scores corresponding to material, organization, expression, and correctness as well as a total score out of 100 were obtained for each essay. The raters were also asked to record the time that it took to rate each essay to the nearest minute in the Excel rating sheet (RQ 7.2).

Once all ratings had been collected, intra-rater reliability was first obtained by comparing the researcher's two sets of total score ratings (RQ 3.2). Cronbach's alpha was calculated by

treating each rating as an item. Next, inter-rater reliability was obtained by comparing the average of the researcher's two sets of total scores to the second raters' set of total scores (RQ 3.3). Again, Cronbach's alpha was calculated by treating each rating as an item. For the 10 essays that were rated by two second raters, an average of the two scores was calculated and used. For the other 40 essays, only one score from one second rater was obtained. Lastly, the two sets of scores, one from the researcher and one from a second rater, were averaged, and the average was used as the final score for each essay. The reliability of the final essay scores was estimated taking into account the combination of the two total scores by treating each of the two sets of total score ratings as an item and using Cronbach's alpha (0.77).

The essays produced by the test takers were further analyzed for source choice, source language integration, and quality of writing (RQ 4.3). Firstly, the researcher performed discourse analysis of the essays for source choice by noting every web source that the test taker referred to or used and coded the type of web source as (a) scholarly, (b) news, (c) organization, (d) Wikipedia, (e) personal, or (f) commercial. Secondly, the researcher performed discourse analysis of the essays for use of source language in-text by visiting each of the web pages that test takers referred to and comparing the language in the essays to the language on the web pages. It was determined whether the test taker had (a) quoted, (b) paraphrased/summarized, or (c) copied the source language. Quoting is defined in this study as copying words from the original source and placing them within quotation marks in the essay. Paraphrasing is defined as using the test taker's own words to borrow an idea from an original source. Although there is no clear cut rule to distinguish paraphrasing from copying, the general rule of thumb of four words was used for this study, that is, if a chain of four or more words from a source appeared in a

participant's essay without quotation marks, it was deemed to be copying. Examples of quoting, paraphrasing, and copying taken from pilot study essay data are provided in Table 3.2.

Table 3.2

Examples of Quoting, Paraphrasing, and Copying

Quoting	Original	A recent study from the University of Oklahoma showed that active video games like Wii boxing or Dance Dance Revolution get kids as active as if they were taking a walk. (from http://www.pediatricsafety.net/2010/06/wii-helps-special-needs-kids-get-exercise/)
	Example	"A recent study from the University of Oklahoma showed that active video games like Wii boxing or Dance Dance Revolution get kids as active as if they were taking a walk."(1*) (Pilot Study Student 2)
Paraphrasing	Original	When the staff at Conlee Elementary School in Las Cruces, N.M., began having students do five minutes of Just Dance, an active video game for Nintendo's Wii, at the start of every school day last year, they noticed a trend: Tardiness went down. When the activity started up again this year, the students cheered and clapped, says physical education teacher Celsa Madrid. (from http://www.usatoday.com/yourlife/fitness/2010-10-11-justdance11_CV_N.htm)
	Example	According to a report in USA TODAY, an elementary school in Las Cru already start using video games in daily teaching, they use a Wii to part of the physical education class. (Pilot Study Student 21)
Copying	Original	The authors concluded that there was "still a generational divide between teachers and students in respect of computer games play". (from http://news.bbc.co.uk/2/hi/technology/5398230.stm)
	Example	The authors concluded that there was "still a generational divide between teachers and students in respect of computer games play".[1] (Pilot Study Student 14)

Other discourse features, namely, length of essay in words and use of in-text citations and list of references were also noted for each essay. Each instance of in-text source use was coded as (a) no in-text citation, (b) signal phrase, or (c) parenthetical or numbered in-text citation. The presence or non-presence of a references list at the end of the text was noted for each essay. There was no second rater for discourse analysis of the essay responses.

Coding of Questionnaire Responses and Interviews

The post-test questionnaire, follow-up questionnaire, post-test semi-structured interviews, and follow-up semi-structured interviews with test takers were used to learn about students' perceptions of various aspects of the test. First of all, the post-test questionnaire provided test-taker perceptions of task administration conditions (RQ 2.1) as well as background information on the test takers, including experience with the internet, writing experience, self-assessment of source-use in writing, English proficiency test scores, and length of stay in the United States. This latter background information was used in part to answer RQs 4.5, and 5.1. Secondly, the follow-up questionnaire provided information about writing test scores, grades received for writing courses beyond English 101C (RQ 5.1), test preparation (RQ 6.1), perspectives on score descriptors (RQ 6.2), and washback effects (RQ 7.1). Questionnaire responses were downloaded from Moodle and entered into an Excel spreadsheet. Descriptive statistics were obtained for the survey items. Thirdly, the audio files of the post-test and follow-up interviews conducted with test takers were transcribed, and content analysis was conducted to identify quotes that have relevance to RQs 2.1, 5.1, 6.1, 6.2, and 7.1. The post-test interviews were also used to confirm and shed further light on the findings about test-taking processes and products (RQ 4.3).

Similarly, the expert and stakeholder interviews with previous and current English 101C and 150 instructors were transcribed and analyzed to identify quotes that are relevant to the topics of domain and sampling (RQs 1.1, 1.2, and 1.3), test specification (RQs 3.1 and 4.1), rating rubric (RQs 2.2 and 4.2), score descriptors (RQ 6.2), and washback effects (RQ 7.1).

Analysis of Artifacts

Syllabi, textbooks, and assignment sheets collected from English 101C and English 150 were used to help answer RQs 1.1, 1.2, 4.1, and 4.2. In the syllabi, the course description, course goals and objectives, and schedule of activities were highlighted. In the textbook, the list of topics was scanned, and the tasks and exercises within the chapters were noted. The assignment sheets were analyzed for the essay type, length, and topic requirements.

Quantitative Data Analysis

Quantitative data analysis was conducted using the final averaged essay scores and quantitative data obtained from coding the screen recordings, essays, and questionnaire responses. Table 3.3 below displays the variables used in the quantitative analysis. Spearman rank-order correlation coefficients were obtained between the essay scores and each of the other continuous variables on ratio, interval, and ordinal scales. A biserial correlation coefficient was obtained by comparing the essay scores with the categorical variable on a nominal scale.

Table 3.3

Variables Used in Quantitative Data Analysis

Variable	**Quantification**	**Measurement Scale**	**Possible Range**
Time on test	Total time spent taking the test	Ratio	33-112 minutes
Time spent on web sources	Total time spent searching for and reading web sources		0-2064 seconds
Length of English learning	Total number of years spent learning English		2-15 years
Essay score	Total score out of 100	Interval	67-97.75
Length of essay including references list	Number of words in essay including the references list		239-760

Length of essay excluding references list	Number of words in essay excluding the references list		226-726
Amount of internet use	Total points from 7 post-test questionnaire items (Everyday=4; Several times a week=3; Once a week=2; On occasion=1)	Ordinal	7-28
Amount of source-based writing experience	Total points from 5 post-test questionnaire items (6-point Likert scale)		5-30
Self-assessment of source-use ability	Points from post-test questionnaire item (6-point Likert scale)		1-6
Performance in post-English 101C course (English 150)	Grades converted into numbers (A=12; A^-=11; B^+=10; B=9; B^-=8; C^+=7)		8-11
Number of web searches	Total count of searches in a search engine or database	Nominal	0-10

For all other questionnaire items that pertain to test takers' perceptions of the test prompt, test administration conditions, and score descriptors, descriptive statistics were obtained, including mean, range, and standard deviation.

Summary and Mapping of Research Questions, Data Collection, and Data Analysis

Summarized and mapped in Table 3.4 below are the research questions, data collection methods, and data analysis methods for each inference and backing in the interpretive argument.

Table 3.4

Summary and Mapping of Research Questions, Data Collection, and Data Analysis

Backing Sought to Support Assumption	Research Question	Data Collection	Data Analysis

Domain Description	1. Domain analysis (Expert consensus on important skills, knowledge, abilities, processes; syllabi and textbooks)	1.1	Expert interviews; syllabi and textbooks	Analysis of expert interview transcripts; analysis of syllabi and textbooks
	2. Domain analysis (Expert consensus on possible assessment tasks; assignments sheets)	1.2	Expert interviews; assignment sheets	Analysis of expert interview transcripts; analysis of assignment sheets
	3. Systematic process of task design and modeling	1.3	Expert interviews (Present the test task I have developed to the experts and ask for comments and feedback on the representativeness and appropriateness of the task for the purpose of the test)	Analysis of expert interview transcripts
Evaluation	4. Multiple task administration conditions were developed, trialed, and revised.	2.1	Trial wording of the prompt, delivery mode of the prompt (Moodle quiz or email), and time limit (90 minutes or 120 minutes) and collect student perceptions through post-test questionnaire and student interviews	Analysis of student perceptions through post-test questionnaire responses and student interview transcripts; pilot study
	5. Systematic rubric development (Develop and revise rubric based on expert consensus of important, relevant criteria)	2.2	Expert interviews	Analysis of expert interview transcripts
	6. Rater training and calibration	2.3	Provision of rater training and calibration before rating of essays	Benchmark essays provided to raters
Generalization	7. Systematic development of test spec for producing parallel tasks	3.1	Obtain expert feedback on test spec and parallel tasks through expert interviews	Analysis of expert interview transcripts
	8. Intra-rater reliability	3.2	One rater rates the same set of essays twice to obtain two sets of ratings of essays.	Calculation of intra-rater reliability using Cronbach's alpha
	9. Inter-rater reliability	3.3	Two or more instructors rate the same set of essays to obtain two or more sets of ratings of essays.	Calculation of inter-rater reliability using Cronbach's alpha

	10. Comparative analysis of test task and instructional tasks in English 101C	4.1	Expert interviews	Analysis of expert interview transcripts
	11. Comparative analysis of test rubric and rubrics used in English 101C	4.2	Expert interviews	Analysis of expert interview transcripts
Explanation	12. Examination of task completion processes (using Camtasia screen recordings) and discourse analysis (essays) support the development of and justification for the task.	4.3	Process data; product data	Analysis of process data (Camtasia screen recordings); discourse analysis of test essays
	13. Comparison studies of group differences	4.4	Post-test questionnaire	Examination of correlations between test performance and English learning, writing experience, and internet use through quantitative analysis of post-test questionnaire responses
Extrapolation	14. Criterion-related evidence	5.1	Post-test questionnaire; follow-up student questionnaire and interviews	Examination of correlations between test performance and students' self-assessment of source-based academic writing ability and students' performance in a post-English 101C course that requires source-based writing assignments
Utilization	15. Examination of perspectives of students on equal opportunity to learn	6.1	Post-test and follow-up student interviews	Analysis of perspectives of students
	16. Score descriptors are provided to students along with their test scores; students' and instructors' perspectives on the usefulness, clarity, and interpretability of the descriptors	6.2	Follow-up student questionnaire and interviews; expert interviews	Analysis of questionnaire and interview data for students' and instructors' perspectives on the usefulness, clarity, and interpretability of the descriptors
Implication	17. Washback studies	7.1	Expert interviews; follow-up student questionnaire and interviews	Analysis of instructor and expert interviews; follow-up student questionnaire and interviews
	18. Controlled rating time and timely distribution of score reports	7.2	Record of the amount of time taken to rate each essay during rating sessions	Average amount of time taken to rate one essay

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents and discusses the results obtained from data analysis. Since the results are going to be used as backing that supports the assumptions under the seven inferences in the interpretive argument, this section will be organized according to the order of the inferences and assumptions in the interpretive argument. The data sources used in the analysis were 48 screen recordings of the test-taking process, 50 test essays, 40 post-test questionnaire responses, 6 post-test test-taker interviews conducted within 4 days of the test, 9 follow-up test-taker interviews conducted 10-15 months after the test, 9 follow-up questionnaire responses, 5 instructor interviews, and artifacts in the form of syllabi, textbooks, and assignment sheets. Overall, the findings indicated that the evidence supports assumptions in the interpretive argument for the use of scores from the web-search-permitted and web-source-based integrated writing test as the final exam scores in English 101C, an academic writing course for international undergraduate students at Iowa State University. However, there were a few assumptions under a few inferences that were only partially supported by the evidence and would thus require further data collection and analysis or revision of the test materials in the future.

Domain Description Inference

The domain description inference is warranted if observations of performance on the integrated writing test reveal relevant skills, knowledge, abilities, and processes taught in English 101C. Backing was sought through (a) domain analysis to identify critical skills, knowledge, abilities, and processes as well as possible assessment tasks and (b) systematic process of task

design and modeling to ensure that an assessment task that requires important skills and is representative of the domain can be simulated.

Domain Analysis (Skills, Knowledge, Abilities, and Processes)

Research question 1.1 aimed to investigate how the domain of web-source-based writing in English 101C and college courses can be defined and described in terms of the important skills, knowledge, abilities, and processes needed in the domain. These were identified through the analysis of syllabi, textbooks, and expert opinion. The expert opinion came from interviews with two English 101C instructors who used the same syllabus as the researcher.

Firstly, the syllabus for English 101C defines the goals of the course as follows:

> The course will help students develop a mature writing style and an ability to integrate ideas, personal experiences, and external sources into their own writing. The course will further emphasize writing as a process and help students learn to improve writing through revision and editing workshops.

An important ability that is highlighted in the goal statement is integrating ideas, experiences and external sources in writing. In addition, an important process emphasized in the course is the improvement of drafts through revision and editing. Furthermore, the specific objectives of the course are outlined in the syllabus as follows:

> Upon completion of this course, students will be able to:
- read challenging texts that reflect important themes and demand critical thinking;
- summarize and critique examples of mature writing styles and techniques;
- revise through multiple drafts to complete successful essays;
- construct coherent essays based on reading, interpreting, analyzing, critiquing, and synthesizing texts;
- adapt the structure, content, and tone of their writing to the knowledge and attitudes of their audience;
- use vivid, concrete language; concise, varied sentences; unified, cohesive paragraphs; gender exclusive English; and a college-level vocabulary; and
- proofread, edit, and correct their final copy for common errors of spelling, punctuation, capitalization, and usage.

The third and seventh objectives repeat one of the course goals, improving drafts through revision and editing, while the fourth objective targets the ability of integrating sources in writing. The first and second objectives point to critical reading skills, while the fifth and sixth objectives refer to various writing skills, from global considerations such as audience awareness and organization to more local issues such as sentence variety, cohesion, and vocabulary. The course schedule outlines daily in-class tasks and activities that help students fulfill the goals and objectives of the course. The activities include reading and discussing readings from the textbook; brief lectures on various aspects of essay writing such as brainstorming, thesis statements, essay organizations, and transitional devices; practice quoting, summarizing, and paraphrasing; exercises on grammar, vocabulary, and mechanics; practice finding and evaluating supporting materials; peer response; and peer editing.

Secondly, the textbook used in English 101C was *A writer's workbook: A writing text with readings* by Smoke (2005). It was adopted for use in all sections of English 101C from Fall 2010 to Spring 2012. The textbook is organized into twelve chapters with each chapter following the same structure of pre-reading discussion questions and vocabulary; one main reading, post-reading discussion questions; prompts for journal writing; an introduction to some structural aspect of essay writing; prompts for a formal writing assignment; an introduction to a technique for getting started; revising practice by analyzing a sample piece of writing; and editing practice on an aspect of grammar and an aspect of mechanics. The structure of each chapter shows that the writing process of drafting, revising, and editing is emphasized. The author of the textbook explains that the focus of the textbook is on English for academic purposes, as the readings are taken from textbooks and research studies and the modes of writing introduced in the chapters are typical of college writing (pp. xvii-xx). In particular, the introductions to structural aspects of

essay writing provide information about various essay modes and their structure. The topics of these introductions include paraphrasing (pp. 85-86), writing a summary (pp. 86-91), and writing a research paper (pp. 189-190). To support the introduction to writing a research paper, a getting started segment on finding sources, including the Internet (pp. 190-193), and a mechanics segment on MLA and APA styles for in-text citations and references list (pp. 202-209) are included. Another "mode" that is introduced in the textbook is writing a persuasive essay under test conditions (pp. 138-141).

Thirdly, the two English 101C instructors were asked to identify the important skills, knowledge, abilities, and processes that are taught in the course and that are required for success in the course. Instructor 1 focused on "paragraph writing and how to link the paragraphs together…more like structure of a paper" as well as "what is plagiarism, how to avoid plagiarism, and…how to search internet for…materials." Instructor 1's focus on coherence, structure, and content is further shown in the following quote:

> I thought many of the, many of my students' writing were pretty empty and not well supported arguments or uh kind of very loose structure in their papers, so I really valued coherent writing style, and also I wished they could um got something from outside instead of just say I think, I believe. So in my class, I tended to look at the content and see whether they have a well-developed storyline or argument. Yeah, as long as the grammar is not that interfering. (Instructor 1)

Instructor 5 explained that "academic writing skills and also some vocabulary knowledge" were mainly taught in the course, along with "some oral communication skills" that were taught informally. Instructor 5 further commented that knowing "how to organize their thoughts in a logical way" would enable students to succeed in the course. The skills of summarizing, paraphrasing, and quoting were also taught in the class:

> [T]here was a required textbook for 101C, and then within that textbook I think there were some subsections that discussed like how to paraphrase, how to summarize, and how to quote. So first I get uh had students read those parts first,

> and then we discussed those, the readings that they did in class. And I also prepared some small classroom activities that um helped them to actually use those skills in their writing. (Instructor 5)

In terms of online language help tools that instructors introduced to their students in class, Instructor 1 introduced the Online Writing Lab (OWL), the Corpus of Contemporary American English (COCA), and Google Fight. The websites were introduced and presented to the students within the course management system: "I even had a page on my Moodle to show students what kind of resources they can go, and I especially introduced the thesauruses to my students" (Instructor 1). Instructor 5, on the other hand, introduced dictionaries at the beginning of the semester as help options: "Usually at the beginning of every semester, I introduce some free online learners' dictionary, like Cambridge or Oxford" (Instructor 5). Furthermore, Instructor 5 was participating in a study on Criterion®, an automated writing evaluation tool, while teaching English 101C, so Criterion® was an important help option used by students in the section. Although using web sources was not particularly focused on in the course, the instructor did introduce how to cite web sources: "especially in my lecture part, I did mention like how to use web sources and then cite web sources in their papers" (Instructor 5).

Both instructors agreed on the importance of source-based writing ability for success at the university. Instructor 1 thought source-based writing ability to be very important and referred to the requirements of English 150 as an example:

> I think it's extremely important, even for undergraduates…I think that's one of the key stuff because in later 150 teaching, I in that class I talked about the online resources and how to incorporate stuff from external world…so they need to go to websites and they need to go to some place and do some reading. (Instructor 1)

Instructor 5 also thought source-based writing ability as important and pointed out the issue of international students who come to the US universities with no prior knowledge of how to properly use sources:

> I think it is very important, especially for international students I guess because a lot of students, I mean I'm not sure about other countries, but at least in my country where I came from, it's not. I'm not sure about these days, but when I was a student, we weren't really aware of the importance of using information from the websites because we didn't really, we weren't really taught that borrowing someone's knowledge or someone's knowledge yeah knowledge properly…[For] those who were not really aware of that, then I think it's even more important to know how to really use web sources in their writing. (Instructor 5)

In summary, since the sections of English 101C shared the same syllabus and textbook, the instructors covered mostly the same content in terms of readings from the textbook, essay assignment types and topics, and process of drafting, revising, and editing. However, individual instructors had freedom in choosing specific in-class tasks and activities, which may have resulted in instructors focusing slightly more or less on different abilities and skills.

Domain Analysis (Possible Assessment Tasks)

Research question 1.2 aimed to identify possible assessment tasks that are representative of the domain of web-source-based writing in English 101C and college courses. Sources for defensible task types were the analysis of assignment sheets as well as expert opinion.

First of all, the assignment sheets identified possible assessment tasks in terms of essay type and task requirements. In the semesters of Fall 2011 and Spring 2012, during which test data were collected, the four major assignments used in most sections of English 101C were all direct assessments of writing, meaning that test takers had to actually write an essay to display their academic writing skills. Four essay types were used, which were personal essay, cause and effect essay, compare and contrast essay, and argumentative essay. For these four essay assignments, students had to write 400-600 words, 400-600 words, 600-800 words, and 700-1000 words, respectively. On each assignment sheet, two topic options were given, along with short descriptions of audience and purpose, planning and drafting, steps to take to finish the

paper, suggested readings from the textbook, elements of a successful essay, evaluation criteria, and additional resources.

In English 150, the essay assignments can vary across sections depending on the syllabus chosen by the individual instructors. In the place-based curriculum adopted by many English 150 instructors, students write a letter and two essays as well as design a brochure or poster. The letter asks students to describe a campus place and explain the place's meaning to the students. The two essays require students to refer to and cite outside sources, particularly those on the web. One of the two essays is a profile of a campus program or organization, while the other is a report and commentary on a campus landscape, building, or art.

Instructors 1 and 5, who used the same general syllabus as the researcher, confirmed during the interviews that they used the same four essay assignments. Instructor 5 added that she used online discussions on Moodle and, in some semesters, vocabulary projects as major assignments. However, none of these assignments required students to use outside sources in their essays.

During the interviews, the two English 101C instructors were additionally asked to identify various direct tests of writing that might prompt test takers to display source-based academic writing skills. Interestingly, Instructors 1 and 5 provided similar comments to each other. Instructors 1 and 5 both first suggested choosing an appropriate topic. Instructor 1 said he would "[i]dentify an interesting…and debatable topic," while Instructor 5 said she would "choose a topic that requires extensive amount of searching information on the web." Then both instructors commented on sources. Instructor 1 said he would "make it explicit that they [students] need to get some help or stuff from outside" and that the students "need to find their, the kind of support from somewhere." On the other hand, Instructor 5 would actually "provide a

list of resources that they [students] can choose from" to "limit the scope of the task." Instructor 1 further commented that he would maybe require five paragraphs as the minimal expectation or at least the components of "introduction, conclusions, and some paragraphs in the middle" so that students would not get lost.

To summarize, the two other English 101C instructors and the researcher used a range of essay types or modes, from personal narrative essays to argumentative essays. English 150, on the other hand, required descriptive and expository essays based on sources. Lastly, the two English 101C instructors had similar ideas when it came to designing a test for source-based academic writing.

Systematic Process of Task Design and Modeling

Research question 1.3 aimed to investigate how much experts thought that the web-search-permitted integrated writing test samples important skills and is representative of the domain of web-source-based writing in English 101C and college courses. Interviews with two English 101C instructors who implemented the same syllabus as the researcher were used to answer this question. The instructors were first asked to look at the test task and to provide suggestions for improvement and modification. Then they were asked (a) whether they thought the test task required important skills taught in English 101C and (b) how representative of the domain of source-based writing in college courses the test task is.

The two instructors made several suggestions for improvement and modification of the test task. First, Instructor 1 suggested adding to the directions some time allotments for stages of the test-taking process:

> I think time management could be an issue for many many students. For myself, I would say, I usually get lost in the internet search because there are too many

> links. We click on this and then you saw any hyperlinks on the navigation bar, and you said, that could be interesting, and then. I think that's kind of uh mm black hole of time. Yeah, so for my from for the instruction part, I would maybe add some suggested time allotment, for example, you may want to spend 30 minutes in the search, and blah blah blah. So that's a way to give students a sense of how much time they need to focus on writing, how much time they can do the search. (Instructor 1)

Instructor 1 also suggested being more explicit in the directions regarding correct source use:

> I think you maybe you should give a warning like uh try to avoid plagiarism or inappropriate use, something….And also…as a reminder it would be good to give students an example in a real text, so students got a sense, oh this is what you mean by in-text citation. (Instructor 1)

Another suggestion from Instructor 1 was providing more context and purpose for writing the essay in the directions:

> For this sample item part, I think it's always helpful to give students the reasons why they have to write this, not just because it's a test….because people write for a purpose. Yeah, if it's just a test, ah. Yeah, people may take it, uh just take it for the test, but still it would be great to have some kind of…why they write and for whom they write. And what kind of rhetorical effect do you expect. (Instructor 1)

Lastly, Instructor 1 commented on the minimum word requirement for the essay and suggested raising it to 500 words:

> Since you give them two hours, and they can also go somewhere for sources, 3[00] is a pretty low requirement to me,…maybe 5[00] is reasonable expectation? Because…when they have more resources, they should have more to say….Also in the major assignments, the last one, if I remember correctly, was about 7 to 800 words. (Instructor 1)

Instructor 5 repeated a suggestion made by Instructor 1 regarding the explicitness of the directions on source use. Instructor 5 commented that unless the students had already been instructed throughout the semester, it would be necessary to explain the word "credible" and also "citation-producing websites."

In response to the interview question of "Do you think the test task requires important skills taught in your course?" both instructors replied positively. Instructor 1 responded, "I think

it matches with the syllabus," although "I think it's very challenging." Instructor 5 also thought that "the specific objectives…on the specification" were ones that are important in English 101C.

Another question that the two instructors were asked was "How representative of the domain of source-based writing in college courses is the test task?" Instructor 1 believed that "this [test task] can be very representative," but at the same time, "it depends on the disciplines":

> I could imagine um in some kind of art courses like psychology or history yeah maybe history, sociology, and people have to write based on what you read. So but I don't know about lab report or some other genres….in my class, I would just ask students to do some very general searches. You do not have to go to very technical or professional websites. In many of my students' writings,…I would just uh give credit for some even newspaper citation or even wiki, so it doesn't matter, but in lab report and some other genres, maybe they need to cite textbook or some other similar stuff. (Instructor 1)

Similarly, Instructor 5 pointed out the issue of variance across disciplines:

> Yeah, I think it is quite um representative at this level. But I mean each student has a different, I mean they major in different disciplines, and I would believe that I'm not really sure about how those will be how different those will be from disciplines to disciplines at the undergraduate level, but I believe especially towards the end of their study, so like when they are in junior or senior, there could be some specific ways to like do source-based writing in each discipline. And in that case, this one would be too general, but since this is an ESL class and you can't really accommodate everyone's needs in one class, so considering that part, I think this is quite representative. But then if you want to go like very specifically, then this would be a little too general. (Instructor 5)

In summary, the two English 101C instructors thought that the test task represented the domain of web-source-based writing in English 101C. However, they had reservations about claiming that the test task represents the much larger domain of web-source-based writing in college courses because they believed that some variation may exist across the disciplines in the ways of doing source-based writing.

Evaluation Inference

The evaluation inference is warranted if observations of performance on the integrated writing test are evaluated to provide observed scores that are reflective of the targeted language abilities. Backing was sought through (a) the development, trialing, and revision of task administrative conditions to ensure that the conditions would be appropriate for providing evidence of the targeted language abilities, (b) systematic rubric development to ensure that the scoring rubric would be appropriate for providing evidence of the targeted language abilities, and (c) rater training and calibration.

Task Administration Conditions

Research question 2.1 aimed to investigate how the test takers felt about the test administration conditions (directions and time limit). To answer the research question, the results from both a pilot study and the current study were used. Students' perceptions of the test administration conditions were obtained through the analysis of post-test questionnaire responses and student interview transcripts.

In the pilot study, participants were given 90 minutes to complete the test. The post-test questionnaire for the pilot study revealed that on a 6-point Likert scale, where 1 indicates strongly disagree and 6 indicates strongly agree, students were positive about the clarity of the directions in the prompt (mean=4.79; standard deviation=1.03, n=20) and the adequacy of the 90-minute time limit (mean=5.21; standard deviation=0.71, n=20). However, although the average time taken by test takers was 75.2 minutes, with a range of 44-89 minutes, two students were taking the test until the very last minute, while six more students went past the 80-minute

mark. One student who spent 73 minutes on the test claimed during the post-test interview that the time given was too short:

> (Key: S=Student; R=Researcher)
> S: It wasn't too hard but you know think the time's like really short for the essay.
> R: Really? You had 90 minutes, so an hour and a half.
> S: Yeah, you really. I think it was a little bit short. So for those four papers I wrote, maybe takes like couple of days, so it's really hard to write it one time. I mean, you need to revise it and add something on it…Yeah, could get more time, I mean, like, give us some time to prepare, like us tell us the topic before, then let us think about it, and write it. Probably better. You need time to find strategies or resources. And the rest of the part is ok. You just need a little bit of time on it. (Pilot Study Student 1)

One of the two students that spent 89 minutes on the test said in the post-test interview that the time limit was enough, but she did have to spend considerable time searching for and evaluating information on the internet:

> S: Time is enough is first. It was a lot of time. And I can search from the internet, and that will never happen in other tests.
> R: So was that helpful or did that take up a lot of time?
> S: It's helpful, but at the same time, it will take a lot of time because I will search something and that we will think which is better. And that will take a lot of time. That's why when I finish, everyone is gone. (Pilot Study Student 22)

Since the university's allotted final exam period is two hours, I decided that two full hours should be given to the students to complete the exam in the present study. Furthermore, the directions were slightly revised to emphasize the necessity of using credible sources and citing the sources properly in two places—both in-text and at the end of the text. The use of help options was also explicitly encouraged (compare two versions of the prompt in Figure 3 and Figure 4).

Topic: Video games in education

Instructions:

1. You will have 90 minutes to write an essay that answers the essay question printed below.
2. Your essay should include an introduction, body, and conclusion. Your essay should also use information from internet sources (citing the sources correctly, e.g., According to Smith (2011)…) and your own insights and experience. Support your opinions and strengthen your main point by using internet sources.
3. You may take notes or plan your essay on the scratch paper provided or in the blank Word document.
4. Aim to write at least 300 words.
5. When you are finished, look over your essay and correct any errors before you turn it in.
6. Your essay will be judged on clarity and overall effectiveness, as well as on content, organization, grammar, vocabulary, and correct use of sources.

Essay question: Should video games be used in elementary schools?

Figure 3. Prompt used in pilot study.

Topic: Video games in education

Instructions:

1. You will have 2 hours to write an essay that answers the essay question printed below.
2. Your essay should include an introduction, body, and conclusion.
3. Your thesis and main ideas must be supported by information from one or more credible internet sources (citing the sources correctly in-text and in a references list) as well as your own insights and experience.
4. You may take notes or plan your essay in the blank Word document. You may also wish to type your essay in Word first and then copy/paste the completed essay into the text box below.
5. You are allowed to use online help options, such as dictionaries, thesauruses, grammar checkers, or citation-producing websites.
6. Aim to write at least 300 words.
7. Your essay will be evaluated on material, organization, expression, correctness, and use of internet sources.

Essay question: Should video games be used in elementary schools?

Figure 4. Revised prompt used in current study.

In the current study, the post-test questionnaire revealed that test takers were overall satisfied with the clarity of the directions in the prompt and the amount of time given (see Table 4.1).

Table 4.1

Post-Test Questionnaire Items for Test Administration Conditions (N=40)

Item	**Median**	**Mode**	**Mean (on Scale 1-6)**	**Standard Deviation**
The directions in the prompt were clear.	6	6	5.3	0.911
The amount of time given (2 hours) was adequate.	6	6	5.5	0.679

Note: 1 – strongly disagree; 6 – strongly agree

Furthermore, the test takers who participated in follow-up interviews perceived the test prompt to be generally clear, as can be inferred from the following comments from eight out of nine follow-up interviews, excluding one student (Student 3) to whom I did not ask the question:

> Mm hmm. Um oh the directions are clear enough. They, yeah, I can understand everything. (Student 6)

> I think it's pretty clear like it introduce the how much time I have and how much words I should write and uh and the grading criteria like grading on the (clarify the) effectiveness and content organization grammar vocabulary and all these things are covered before the exam in the practice (since last) in the assignments before so pretty good. And also and also (?) ask me to open up a blank Word document. That's very. Yeah, that's considerate. Yeah. (Student 22)

> S: Yes, I think it's quite clear and I and this is your put you told me I was should use what and what so I know what I should include in my essay. And requirement, yeah, it's quite clear.
> R: Was it clear that you had to search the internet for sources and use the information?
> S: And my own insights and experiences, that means not all is (and?).
> R: Right, so you have to use both. Yeah.
> S: Have both, yeah.
> R: Both your opinions and the source.
> S: Yeah, and you have the words requirement. So it's quite clear. (Student 30)

> R: Um in the directions, I included information here, directions here, that the students have to use at least one outside source in their essays, so was that point clear within the directions?
> S: I think you had highlight this, yeah, so, yeah it's clear. So I don't, sorry I don't think it's a question, so it's pretty clear. (Student 39)

> I think it's enough. The instruction is enough for us to understand. (Student 42)

> S: I think it is clear.
> R: Are there any confusing or unclear places? Or ambiguous.
> S: No. (Student 45)

> It is pretty clear because the instruction wrote down all the requirement and all the details and yeah. So what you should do and what you shouldn't do. So like you have to write at least 300 words. And you are, you must like get some information from one or more reliable internet source. And what assign, what the thesis of the paper is, mm hmm. It's very clear. (Student 47)

> Yeah, I think it's very clear because first say how, your limit in time. And then the introduction, body, and conclusion. The whole structure is very clear. It gives lots of advice how to write it better like you can use the help options dictionary and citation-producing website and grammar-checkers because before this exam, you give us lots of website to grammar-checkers and citation producing websites, so it's very helpful and we can also use this source to help our exam. And it's how much how many words we need to write. So it's very clear, I think. (Student 48)

The average amount of time taken on the test was 71 minutes with a range of 33-113 minutes for the total of 50 test takers. The average length of time was slightly shorter than the average time taken by test takers in the pilot test, which was around 75 minutes, while the range was much wider in the current study.

Of the forty students who completed the post-test questionnaire, four students commented on the adequacy of the time limit in their response to the open-ended question asking what they liked about the test:

> It was challenging but not impossible. These kind of essays are good for tests because they do not require much time to write the required length. (Student 17)

> The time is adequate. (Student 26)

> [C]lear and enough time. (Student 34)

The test gives sufficient time to make our essay as good as possible. (Student 39)

Another student expressed a similar opinion during the post-test interview:

R: Did you think you had enough time?
S: Yes, two hours' time is very, yeah it's very, it's enough for me. (Student 29)

A new time-related issue with test administration conditions that arose in the current study was the time at which the test began. For one of the three sections in the study, the test began at 7:30 am, which was much earlier than the 9:00 am at which the section had been meeting for classes during the semester. Two students from that section had the following to say in the post-test questionnaire about what they disliked about the test: "The test time is too early and the weather is not good" (Student 11); "Time is so early that I feel sleepy" (Student 15). No similar comments were made by students in the other two sections that began the test at 9:30am. It is possible that the time at which the test was given might have affected some test takers' performance. However, as the university decides on the final exam schedule before the beginning of each semester, instructors have no control over what time of day the exam can be given.

To summarize, based on the post-test questionnaire responses, post-test interviews, and follow-up interviews, test takers mostly perceived the directions to be clear and the time limit to be adequate in the current study.

Systematic Rubric Development

Research question 2.2 investigated what experts thought about the appropriateness of the rating rubric for providing evidence of web-source-based academic writing ability. Transcripts of interviews with five previous and current English 101C instructors were analyzed to find the

answer to this question. During the interviews, instructors were first asked to share their ideas of what criteria or aspects should appear in a rubric that scores essays from a test of source-based academic writing. This question was asked before the instructors had a chance to look at the rating rubric. Then the instructors were given a copy of a previous version of the rating rubric (Appendix J) to read and were asked to provide comments about the rubric for revision and improvement.

I read through the transcripts to create a summary of expert opinion on what criteria or aspects should appear in a rubric that rates essays from a test of source-based academic writing. Classified and summarized in Table 4.2 are the comments that the five English 101C instructors provided.

Table 4.2

Expert Opinion on Criteria to be Included in a Rubric for a Test of Source-Based Academic Writing

Criteria	**Instructor 1**	**Instructor 2**	**Instructor 3**	**Instructor 4**	**Instructor 5**	**Total**
Reliability and usefulness of sources	X					1
Correct use of sources (citation)	X		X	X	X	4
Integration of source information	X	X "synthesis"		X	X	4
Accuracy of information from sources		X "paraphra-sing, summari-zing, quoting"			X	2
General criteria that are usually used in			X "very basic things about writing,	X "content, organiza-tion, logical	X "general criteria that are usually	3

academic writing			such as you know structuring, content, yeah, language"	thinking, critical thinking, delivery, language, syntax, grammar... mechanics"	used in academic writing in general"	
Grammar	X					1

Most instructors agreed on the necessity of including integration of information from sources with the writer's own words and correct source use in the form of in-text citations and end-of-text list of references, in addition to other general criteria that would appear in a rubric for an academic essay, such as content, organization, and language. The previous version of the rating rubric already included all of these aspects that were mentioned by the instructors.

Table 4.3 summarizes and classifies the five instructors' suggestions for revision and improvement of a previous version of the rating rubric. These suggestions were given by the instructors after they had been provided with a copy of the previous version of the rating rubric and were given time to read it.

Table 4.3

Instructors' Suggestions for Revision and Improvement of Rating Rubric

Suggestions	**Instructor 1**	**Instructor 2**	**Instructor 3**	**Instructor 4**	**Instructor 5**	**Total**
Raise the bar on plagiarism	X any plagiarism should get lowest level			X		2
Clearer distinction between levels (stricter standards and higher	X excellent should have something more than the basics			X "Excellent should be like really outstanding"		2

expectations)					
Give "use of internet sources" a separate row as a criterion	X			X "a separate criteria for the um using resources part"	2
Combine expression and correctness		X "language"	X "style"		2
Create a formatting criterion		X include citation under "formatting"	X include citation under "delivery" with formatting		2
Add a new criterion for integration			X		1
More clarification, elaboration, or specification of terms	X essay length, thesis statement, transitions, cohesive		X specify scope under material, e.g., "fully developed"	X "a little more elaborations, especially for the lower levels"	3
Have a score range within each level	X				1
No levels		X allow instructors to give any score within a range for each criterion			1

The first two suggestions are to make the levels within the rubric stricter so that the excellent level has higher expectations in terms of all criteria, and plagiarism is a cause for major

downgrading. Instructor 1, for example, thought that basic requirements should come toward the lower end of the levels, while higher expectations should be reflected in the higher levels:

> (Key: I=Instructor; R=Researcher)
> I: So at a higher level you can expect of course some performance like this [basic performance], but you have higher uh better performance on a higher level.
> R: So sort of building up from the lowest level and adding on elements?
> I: Yeah because in my class when I used the rubric in my 150 class, I asked my students to read from the lower level because I believe psychologically when you read the high level, excellent, and you may implicitly map your performance to that one. And ah! Yeah, pretty good. That is ok. And then you got maybe a wrong judgment. But if you look at the low level, and you got more careful, and so ah, should, am I like this one? So I think mentally maybe there could be some different effects. (Instructor 1)

Similarly, Instructor 4 pointed out the higher expectations linked to the excellent level and also the belief that plagiarism should be more severely downgraded:

> Excellent should be like really outstanding, and good you know should be ok but not contain instances of serious error. Uh huh. Yeah, so for the student, if there are covert plagiarism and attempted you know isolated copying and still the student get 26 point out of that, and that means good, then the student is not gonna depart from that habit. It's not you know discouraging enough. So this rather encourages that….It's not deterring enough. (Instructor 4)

The next four suggestions relate to the issue of what criteria to include within the rubric, particularly with regard to where to place source use within the rubric. Instructors 1 and 5 thought that a fifth criterion should be added to the rubric to deal with all aspects of source use, including the integration of information from outside sources with the writer's own words and the correct citation of sources. Instructor 1 suggested creating and comparing two versions of the rating rubric, one version with four criteria—Material, Organization, Expression, and Correctness—and another version with a fifth criterion of Source Use added:

> If you have two rating rubrics, it will be interesting to look at the uh rater performance on these categories. And sometimes if my guess is that since the title here material and correctness are so uh kind of uh give the sense of grammar correctness and content completeness or something, so they may kind of cover

> some of the parts in your sources related part. That could be could obscure some of the your intended part. (Instructor 1)

Instructors 3 and 4 thought that expression and correctness should be combined into one criterion of "language" or "style" and that a new criterion of "formatting" or "delivery" should be added to the rubric. The issue of correct citation of sources would be subsumed under the new criterion. Instructor 3 clarified what specific aspects would go under the two criteria:

> I would put everything else into grammar or into language, including expression and word forms and structure, and then the citation and like word font, you know consistent for word type, word font, and uh double-spacing, things like that would be categorize those along with the citation as formatting. (Instructor 3)

Instructor 4 further suggested that a new criterion be added for integration of source material, resulting in five criteria in total—material, organization, style, delivery, and integration.

Three instructors commented that several terms in the rubric should be clarified and specified. This would make rating clearer for instructors because they would know exactly what to be looking for as they rate the essays. Instructor 2, for example, provided specific details that should accompany several terms in the rubric, including thesis statement, transitions, and cohesive:

> I think there are two elements that must be graded in the thesis statement. There is the topic and controlling idea. So in one thesis statement you must check for both separately….Transitions again needs to be opened up again I think at this level because you're teaching final now. I think word patterns we need to name this transitions now. What kind of transitions do we mean? Do we mean, I mean, conjunctions here? Do we mean, I mean, um the demonstrative pronouns? Word patterns, for example. Or do we mean reporting verbs? Now there are a lot of noun families that can be under this umbrella….Cohesive can be treated in two ways, I think. One between the paragraphs and one…within the paragraphs, one body and thesis. Two ways to establish that [cohesion]. They might have one and not the other one. So they might have cohesive in the paragraph but not, the paragraph might not be connected to the thesis statement. (Instructor 2)

At the same time, students would also benefit from the elaborations, as they can more easily identify their weaknesses and think about how to improve those weaknesses. Instructor 5 commented specifically on how the elaborations can benefit the students:

> Maybe you could provide a little more information especially when some flaws are mentioned. So for example here for this part, I mean yeah development is insufficient, maybe that's transparent, but examples maybe inappropriate, in what way? Yeah. Those kind of information might be really helpful especially for students cause this is just telling them, ok you're wrong, but then they will think, ok so I'm wrong, but how can I be improved? And that kind of information is missing currently here in this descriptor. Yeah, so if you consider creating a separate version for students, then you probably need add some more information especially this kind of part, for educational purpose...for the some parts that they need to improve on, they want to know how, so what's wrong?…if you can explain a little more about what's wrong, then they could easily um make a plan on how they can improve those parts. (Instructor 5)

The last two comments appeared to reflect Instructors 2 and 3's preferences for assigning scores. Instructor 2 preferred to have a score range within each level instead of just one score per level, while Instructor 3 did not want levels at all but rather wished for a range within each criterion so that a score can be determined after consideration of various requirements for a criterion.

The rating rubric was revised based on the comments and feedback from the five English 101C instructors to be used for the rating sessions (Appendix K). Firstly, the rubric was revised to include stricter standards regarding plagiarism and correct use of sources. Secondly, several terms in the rubric were clarified. Thirdly, I decided to keep the original four criteria (Material, Organization, Expression, and Correctness) because I considered consistency to be important in the test use context. Since the four criteria had been used to grade the four writing assignments in the course sections, I thought that using the same criteria again in the final exam would create less confusion for students. Also, there was a mix of ideas from the instructors regarding this issue of criteria to include in the rubric, and it was difficult to take into account all of the

suggestions at the same time. Lastly, I decided to create a separate student version of the rubric with more simplification and elaboration of terms.

Rater Training and Calibration

Research question 2.3 was "How much can instructors be trained to avoid bias for or against different groups of students?" The second raters in this study were sent the rubric and four benchmark essays to self-train before beginning to rate their ten assigned essays. The essays had been stripped of any identifying information. Because of this blind rating procedure, I believe that bias was avoided to the extent possible. There was no significant difference in mean essay scores between male students and female students ($t=0.992$, $df=48$, $p=0.326$, mean difference=2.371, n[male]=33, n[female]=17). There was also no significant difference in mean essay scores for Chinese students versus non-Chinese students ($t=1.133$, $df=48$, $p=0.263$, mean difference 3.689, n[Chinese]=43, n[non-Chinese]=7).

Generalization Inference

The generalization inference is warranted if observed scores are estimates of expected scores over the relevant parallel versions of the tasks and within and across raters. Backing was sought through (a) systematic development of test specifications for producing parallel tasks, (b) estimation of intra-rater reliability, and (c) estimation of inter-rater reliability.

Systematic Development of Test Specification for Producing Parallel Tasks

Research question 3.1 was how much experts found the test task specification to be well defined for producing parallel tasks. To find the answer, five instructor interview transcripts

were analyzed. In response to my interview question of whether they would be able to create parallel tasks based on the test specification, all five English 101C instructors answered yes.

Specifically, one instructor said that creating parallel tasks based on the specification is "doable" (Instructor 1), while another said "I'll be able to create a test based on this, the information given in this specification" (Instructor 5). Instructor 3 commented that creating parallel prompts might be easy because the test allows test takers to search the internet for information:

> I think so. I think it's easy. Yeah, it's just like you know coming up with prompts of the same difficulty level. And I don't think it's a big deal, um no matter which task you give them. I think the internet is pretty powerful. They can you know no matter which prompt they get, they can, they won't, it won't be a very hard job for them to do the search. So I think yeah, it's pretty easy, I think. (Instructor 3)

Instructor 2 suggested that parallel tasks can be created by borrowing from Criterion® prompts and by building an item bank. The instructor also raised the important issue of creating multiple prompts every semester, so that each section of English 101C would be given a different prompt and so that there would be no issues with students exchanging information about topics across sections:

> I: Yeah, I mean there are tons of prompts that we take from Criterion®, but we have also a bank that we have you know from I don't know how many years, so why not? And we have to, we change those prompts every semester because we have to do it. Because now there are a lot of people coming from the same country and then they have a good unity here, social unity, so say if one student were in your class last semester, and might be in my class, you know her friend, roommate, I say, so we have to do we have to be very careful. We have to change, recycle this, but that means that we have to produce a lot of items.
> R: So producing items you don't think is an issue.
> I: I mean, yeah, as long as it is we know the genre, we know the task, we know what we're asking. We need like at least 10 of them per semester because there are 10 sections at least. Every section must get a different one. That's another thing. So yeah. So 10 at least per semester. But then after four, I say after the 4th semester, you can recycle them, which is gonna be safe, hopefully. (Instructor 2)

Although the five instructors thought that they would be able to create parallel tasks based on the test specification, it should be acknowledged that it may be more difficult than these responses suggest to actually create the parallel tasks (from an assessment perspective). Thus, it is necessary to investigate this question more fully by actually asking instructors to create tasks from the specification and comparing the tasks to see how parallel they are in terms of criteria such as difficulty and availability of web sources.

Intra-Rater Reliability

Research question 3.2 was "How high is the intra-rater reliability?" Table 4.4 below displays the descriptive statistics for the researcher's two sets of ratings with a six-week period in between. The measure used to estimate intra-rater reliability was Cronbach's alpha, which was calculated by treating each rating as an item. The Cronbach's alpha was 0.885, which indicates a good level of reliability.

Table 4.4

Descriptive Statistics for Researcher's Two Sets of Ratings

	Mean	**Standard Deviation**	**N**
Rating 1	83.00	7.134	50
Rating 2	83.76	7.018	50

Inter-Rater Reliability

Research question 3.3 was "How high is the inter-rater reliability?" Table 4.5 below displays the descriptive statistics for the researcher's averaged set of ratings and the second raters' set of ratings. The measure used to estimate inter-rater reliability was Cronbach's alpha,

which was calculated by treating each rating as an item. The Cronbach's alpha was 0.770, which indicates an acceptable level of reliability.

Table 4.5

Descriptive Statistics for Researcher's and Second Raters' Sets of Ratings

	Mean	**Standard Deviation**	**N**
Researcher	83.38	6.70254	50
Second rater	82.75	10.62402	50

For the generalization inference, the three types of backing supported the three assumptions under the inference. However, more concrete research on parallel task production based on the test specification would strengthen the backing for the first assumption under the inference.

Explanation Inference

The explanation inference is warranted if expected scores are attributed to a construct of web-source-based academic writing ability, which is defined by the English 101C syllabus and the teaching/learning activities in the class. In other words, the construct measured on the test is the intended construct of web-researching-to-write. The first type of backing sought for the explanation inference is similarity of the test task and test rubric to the instructional tasks and rubrics used in English 101C. Because the instructional tasks and rubrics used in English 101C are intended to develop and elicit the same construct of web-researching-to-write, these comparisons provide evidence for the use of the test task and test rubric by showing that the test task and test rubric indeed elicit and measure the construct of web-researching-to-write. The comparisons were made on the basis of instructors' judgments. Further backing was sought through an examination of task completion processes and discourse analysis of essays to ensure

that the processes and essays reflect the intended test construct of web-researching-to-write and thus support the development of and justification for the test task. Finally, backing was sought through a comparison of group differences by examining relationships between students' test performance and their English learning, writing experience, and internet use. This evidence speaks to the web-researching-to-write construct that the test is intended to measure by showing that the test scores reflect different aspects of the construct.

Comparative Analysis of Test Task and Instructional Tasks

Research question 4.1 asked how much the test task requires the abilities taught in the instructional tasks in English 101C. To find the answer to this question, the perceptions of experts were obtained through interviews with two English 101C instructors who shared the same syllabus with the researcher. Both instructors thought that the test task reflected the instructional tasks in English 101C in many ways.

When asked "To what extent does the test task reflect the instructional tasks that are used in your course?" Instructor 1 responded: "Yeah, I think that's relevant and because in my class I did expect students to do some search on their own. Searching. Search again research….Yeah, so I think that's, that makes sense." Instructor 1 included paraphrasing, taking notes on a text, analyzing one-paragraph and one-sentence summaries, and punctuating direct speech in week 11 of his weekly plan, while the topics of library and internet research, quoting, and citing and documenting sources were on the agenda for weeks 13 and 14. Instructor 1 also mentioned introducing numerous online resources for writing to his students, including COCA, Google Fight, and thesauruses.

To the same interview question, Instructor 5 responded by comparing the requirements of the test task to the course content, course assignments, and final exam. The similarities pointed out by Instructor 5 to exist between the test task and course content were (a) writing of a "multi-paragraph essay…That's what I have had emphasized every time almost every class to my students"; (b) "online help options" including dictionaries and Criterion®; (c) "organization"; and (d) "development of paragraphs, that was the most important thing that my students would know by the end of the semester. And then yeah details, examples." Instructor 5 also noted that the test task resembled the final exam used at the end of her class, which was a timed argumentative essay writing exam:

> Especially for the final exam, I also did a timed writing each semester, but it wasn't really, oh it was argumentative essay actually, but then, and then they were allowed to use dictionary and also Criterion®, but not additional materials for the contents of their essays. Yeah. So it has to be it had to be all based on their personal experiences….So this [test task] looks quite similar to my timed final exam, but except choice of using web-based resources in their writing. (Instructor 5)

Although the final exam in Instructor 5's class did not require or allow students to refer to web sources in their essays, the major writing assignments did allow the use of sources:

> But for their actual major papers, not final exam, they were allowed to use sources. Actually they were required to use at least one, yeah, in their paper….but in my actual class, I didn't really specify any specific sources to use in their in students' writing, so in that sense yours [test task] is closer to what I actually did in my 101C. (Instructor 5)

Comparative Analysis of Test Rubric and Course Rubrics

Research question 4.2 asked how much the test rubric reflects the rubrics used to evaluate writing in English 101C. To find the answer to this question, the perceptions of experts were obtained through interviews with two English 101C instructors who shared the same syllabus as

the researcher. Both instructors thought that there was much overlap between the test rubric and rubrics used in English 101C.

The two English 101C instructors were asked the question, "To what extent does the rating rubric reflect the rubrics that are used in your course?" Instructor 1 responded that the test rubric reflected the same four criteria that were in the rubrics used in the course, though the test rubric was much more detailed than the English 101C rubric:

> I think it's pretty good. Much better than the one I used. Mine is uh I should say ours was very simple. Yeah. Because we only had, um we had the general categories and very limited information of each category. This is more like 150, yeah…the one I used was too simple. (Instructor 1)

Instructor 1 also noticed that the test rubric resembles the rubrics used in English 150, which were very detailed. Instructor 5 thought that the test rubric reflected the rubric used in English 101C, though she recalled adding a fifth criterion for source use to the rubric used to grade an essay assignment that required the use of outside sources:

> I think overall, this looks very similar to the rubric that I used for my 101C, except I think I had a separate criteria for the um using resources part for the major paper….Cause that was a new thing that was introduced in that unit and I probably wanted to highlight the importance of that part, so I had a separate criterion only for that. But other than that, this looks very similar to the rubric that I used yeah. (Instructor 5)

Test Completion Processes and Discourse Analysis of Products

Research question 4.3 was "What test-taking processes did test takers follow, and what web-searching behaviors did test takers show? What online language help tools did test takers consult? What relationships are there between test-taking processes and test scores? Do the test scores reflect how well web sources are used in the essays? How do the selection of sources, attribution to sources, and integration of source language relate to scores or differ across score

levels?" These questions were answered based on analysis of 48 screen recordings as process data and discourse analysis of 50 test essays as product data. These data were analyzed quantitatively to find correlations between the test score and the other quantitative variables that represent test-taking processes and quality of writing. For certain quantitative variables that are not amenable to correlations with test scores, cross-level comparisons were made after grouping the essays into score levels.

Test-Taking Process

Camtasia screen recordings of 48 students were coded for test-taking activities. Recordings of two students were lost and could not be included in the analysis. The times spent by the 48 test takers on various test-taking activities during the test session are displayed in Table 4.6.

Table 4.6

Time Spent on Test-Taking Activities (N=48)

Activity	**Mean**	**Standard Deviation**	**Range**	**Median**
1. Read prompt	0:03:12	0:02:01	0:00:50 - 0:10:36	0:02:39
2. Write, edit, proofread	0:42:38	0:15:18	0:14:07 - 1:18:09	0:41:15
3. Think, read essay	0:03:33	0:03:03	0:00:00 - 0:13:04	0:02:53
4. Search, click on links	0:02:30	0:02:02	0:00:00 - 0:07:43	0:02:13
5. Read web source	0:07:45	0:06:26	0:00:00 - 0:31:35	0:06:35
6. Help option	0:05:22	0:04:50	0:00:21 - 0:17:54	0:03:27
7. Use source (Copy, paste, citation, references list)	0:01:41	0:02:00	0:00:00 - 0:10:31	0:01:03
8. Other (e.g., adjust windows,	0:03:33	0:02:42	0:00:01 - 0:13:32	0:02:53

save document)				
Total	1:10:14	0:21:53	0:32:42 - 1:52:49	1:05:43

There was no significant correlation between total time spent on test and essay score (Spearman's rho=0.045, p=0.755, n=50). There was also no significant correlation between time spent on web sources (activities 4 and 5) and essay score (Spearman's rho=-0.033, p=0.826, n=48).

Table 4.7 summarizes the search engines used by the test takers to search for sources on the internet along with the number of web searches conducted using each search engine.

Table 4.7

Search Engines and Databases Used for Web Searches

Type	**Search Engine**	**Number of Test Takers**	**Number of Searches**
Commercial	Google	36	99
	ISU Google	2	5
	Wikipedia	2	4
	Baidu	1	2
Academic	Library Quick Search	2	3
	Academic Search Premier	1	2
	EBSCOhost	1	2
	ERIC	1	1
	Google Scholar	1	1
Total			119

Test takers did an average of 2.5 searches for web sources. Google was the most popular commercial search engine used by 75% of the test takers at least once. Five students used one or more academic search engines or databases through the library website. Five students conducted no web searches at all. There was no correlation between the number of searches conducted and the essay score (Spearman's rho=-0.018, p=0.902, n=48).

Table 4.8 lists all content words that were included three or more times in the search key words or phrases, after correcting spelling mistakes or typos and combining singular and plural forms. Articles and prepositions were excluded from the list. The raw frequency list, which was obtained using Web Frequency Indexer v. 1.3 (http://www.lextutor.ca/freq/eng/), is provided in Appendix O.

Table 4.8

Content Words Frequently Included in Search Key Words and Phrases

Content Words Included in Search Key Word/phrase	**Number of Key Words Out of 119**
game(s)	104
video	104
elementary	42
school(s)	39
used	23
be	22
education	20
should	14
teaching	8
children	7
effect(s)	5
how	5
violence	4
what	4
can	3
disadvantage(s)	3
educational	3
good	3
important	3
kids	3
play	3
student(s)	3

Test takers tended to select search words from the test prompt, which was “Should video games be used in elementary schools?” Also used in search key phrases were words semantically related to words in the prompt, such as education(al) and teaching to imply the meaning of

"using in schools." Some test takers chose to use key words that portray a pro or con viewpoint on the issue, such as good and important for the pro side and violence and disadvantage for the con side.

Table 4.9 summarizes the online language help option types that test takers consulted during the test along with the number of uses of each help option. Only help options that are on the internet are included in the table. This means that Microsoft Word's grammar checker, spell checker, autocorrect, and word count and Mac's spell checker have been excluded. One section of the course included test takers who composed their essays entirely in the text box of a reply email because that section was emailed the prompt and had to submit their essays by email. Some students in the other two sections composed their essays in the text box within the Moodle quiz. For these two groups of students, only the Mac spell checker was functioning within the text box, so they did not have access to the grammar checker, autocorrect, and word count functions, which are only available in Microsoft Word. Therefore, a spell checker was available to and used by all 48 test takers, while the other three Word-proprietary help options were not.

Table 4.9

Online Language Help Options Consulted During Test

Help Option Type	**Help Option**	**Number of Test Takers**	**Number of Uses**
Dictionary/ translator	Youdao	8	141
	Google Translate	9	68
	Baidu	13	36
	dictionary.com	3	18
	Naver Dictionary	1	13
	dict.cn	2	9
	iciba.com	1	9
	Google (as dictionary/corpus)	5	9
	Longman English Dictionary Online	1	6
Phraseology	Student's 4th essay attached to email (to check phraseology)	1	2
Grammar/spell	www.paperrater.com	3	3

checker	www.grammarly.com	2	2
	www.grammarcheck.net	1	1
	www.spellchecker.net/grammar	1	1
	www2.elc.polyu.edu.hk/CILL/errordetector.htm?mid=544	1	1
	www.gingersoftware.com/grammarcheck	1	1
Citation	APA Citation Style PowerPoint slides (English 101C material available in Moodle)	6	30
	Citation Machine	6	12
	KnightCite	1	1
	Citation Maker (Oregon Public Education Network)	1	1
	A Word document attached to an email with references list (used as template)	1	1
Word count	http://www.mb5u.com/tool/zishutongji	1	1
	http://tool.sougee.com/word-counter	1	1
	http://www.ateste.com	1	1
	http://www.4808.com/paiban2.asp	1	1
Total			369

The most popular type of online language help option was dictionaries and translators. More students used dictionaries in their L2 or Google Translate to translate words back and forth between English and their L2, although some students used English-English dictionaries or Google in English. Dictionaries and translators were followed by resources for creating in-text citations and references list entries. Several students either used the two online grammar checkers that the instructor had introduced during the semester or searched for new ones. One student searched for word count programs online because he failed to find the word count function in Microsoft Word.

Test Product

The average length of essays was 401.14 words excluding the references list (range: 226-726) and 413.64 words including the references list (range: 239-760). There was no significant

correlation between essay score and length of essay, excluding references list (Spearman's rho=0.196, p=0.173, n=50) or including references list (Spearman's rho=0.212, p=0.140, n=50).

To investigate whether the use of web sources in the essays varies with performance, I decided to divide the test takers into groups to allow for certain comparisons across levels. Before doing this, I needed to make sure that the three sections of test takers can be considered as one homogeneous group. There was no significant difference between the test score means for the three sections of students according to a one-way ANOVA (F(2, 47)=1.319, p=0.277). Therefore, all 50 test takers were considered as one group. The 50 students were then divided into three new groups based on essay score. The first grouping was done based on the total essay score, with 17, 17, and 16 students in each group. The second grouping was done based on the material component score, with 17, 16, and 17 students in each group. The third grouping was based on the correctness component score, with 14, 22, and 14 students in each group. The main consideration when dividing the essays into groups was major breaks in the score distribution, with a secondary goal of having equal group sizes. The descriptive statistics for the three groupings are displayed in Tables 4.10, 4.11, and 4.12.

Table 4.10

Descriptive Statistics for Essays Grouped According to Total Essay Score

Group_Total	**N**	**Mean Score**	**Standard Deviation**	**Range**
1 (High)	17	92.4265	2.82257	89.5 - 97.75
2 (Mid)	17	81.9706	3.03223	78.25 - 86.25
3 (Low)	16	74.2813	2.86047	67 - 77.5
	50	83.0650	8.00762	67 - 97.75

Table 4.11

Descriptive Statistics for Essays Grouped According to Material Component Score

Group_M	N	Mean Score	Standard Deviation	Range
1 (High)	17	28.3088	1.1475	27 - 30
2 (Mid)	16	24.9844	0.8778	23.5 - 26
3 (Low)	17	21.6618	1.1955	19 - 23
	50	24.985	2.966	19 - 30

Table 4.12

Descriptive Statistics for Essays Grouped According to Correctness Component Score

Group_C	N	Mean Score	Standard Deviation	Range
1 (High)	14	18.6964	0.6292	17.75 - 20
2 (Mid)	22	15.9318	0.8970	14.5 - 17
3 (Low)	14	13.4286	0.7871	11.75 - 14
	50	16.005	2.1408	11.75 - 20

First of all, the web sources used in the essays were considered. Regarding the number of sources used, a total of 81 web sources were used by 41 students. One source used by two different students was counted as having been used twice. Also, one source used by one student in two different places within the essay text was counted as having been used twice. Nine students did not use any web sources in their essays. There was no significant correlation between the essay score and the number of sources used in the essay (Spearman's rho=-0.058, p=0.688, n=50).

The types of web sources were also considered. Credible sources included scholarly articles or web pages, news articles, and organizational web pages. Non-credible sources included Wikipedia entries, personal web pages or blogs, and commercial web pages. Table 4.13 displays the types of web sources used in the essays, with essays grouped according to total essay score, while Table 4.14 displays the sources used, with essays grouped according to the material component score.

Table 4.13

Types of Web Sources Used in Essays (Grouped According to Total Essay Score)

Group_ Total	**Credible**			**Non-Credible**			**Total**
	Scholarly	**News**	**Organization**	**Wikipedia**	**Personal**	**Commercial**	
1 (High) n=17	1	12	1	3	4	6	27
2 (Mid) n=17	10	9	2	2	3	3	29
3 (Low) n=16	3	9	1	5	3	4	25
N=50	14	30	4	10	10	13	81

Table 4.14

Types of Web Sources Used in Essays (Grouped According to Material Component Score)

Group_ M	**Credible**			**Non-Credible**			**Total**
	Scholarly	**News**	**Organization**	**Wikipedia**	**Personal**	**Commercial**	
1 (High) n=17	1	10	2	3	4	5	25
2 (Mid) n=16	7	15	1	2	2	2	29
3 (Low) n=17	6	5	1	5	4	6	27
N=50	14	30	4	10	10	13	81

It seems that neither the total essay scores nor the material component scores reflect the type of sources used in the essays in terms of credibility because the essays that received higher total essay scores or material component scores did not use more credible sources than essays that received lower scores, that is, the numbers in the upper left-hand corner were not higher than the numbers in the lower left-hand corner in both Table 4.13 and Table 4.14. Similarly, essays that received higher total essay scores or material component scores did not use fewer non-credible sources than essays that received lower scores, as can be inferred from the fact that the numbers

in the upper right-hand corner are not lower than the numbers in the lower right-hand corner in both Table 4.13 and Table 4.14. In fact, essays belonging to the middle group in both groupings used the most credible sources and the fewest non-credible sources.

The second consideration was attribution to sources in a list of references at the end of the essay. An independent samples t-test between the mean essay score for the group of students who included a references list in their essays and the mean essay score for the group of students who did not include one showed a significant difference at $\alpha=0.05$ level ($t=2.8$, $df=48$, $p=0.007$, mean difference=5.945, 95% C.I.=[1.675, 10.214]). The mean material component scores were also significantly different between the two groups at $\alpha=0.05$ level ($t=2.845$, $df=48$, $p=0.007$, 95% C.I.=[0.6546, 3.8101]), as were the mean correctness scores ($t=2.066$, $df=48$, $p=0.44$, 95% C.I.=[0.0327, 2.3920]). Table 4.15 displays the descriptive statistics for the two groups.

Table 4.15

Descriptive Statistics for Groups According to Presence of a References List

	Mean Essay Score (Standard Deviation)	**Mean Material Score (Standard Deviation)**	**Mean Correctness Score (Standard Deviation)**	**N**
References list	86.1563 (7.437)	26.1458 (2.6763)	16.6354 (2.2128)	24
No references list	80.2115 (7.560)	23.9135 (2.8674)	15.4231 (1.9349)	26

The third consideration was attribution to sources in-text. An independent samples t-test between the mean essay score for the group of students who included one or more in-text citations (signal phrase, parenthetical citation, or numerical citation) in their essays and the mean essay score for the group of students who did not include any did not show a significant difference ($t=1.037$, $df=48$, $p=0.305$). Table 4.16 displays the descriptive statistics.

Table 4.16

Descriptive Statistics for Groups According to Presence of In-Text Citations

	Mean Essay Score	**Standard Deviation**	**N**
In-text citation(s)	83.9453	8.001	32
No in-text citations	81.5	8.002	18

The next step in the discourse analysis was to consider the use and integration of source language in-text in tandem with the attribution to sources in-text. The integration of source language in-text was categorized into copying and correct use, with the latter category including quoting, paraphrasing, and summarizing. Attribution to sources in-text was categorized into no in-text citation and some form of in-text citation, with the latter category including signal phrase, parenthetical in-text citation, and numbered in-text citation. I excluded from analysis three cases where a source was listed in the references list but not used in-text and two cases where a source was just introduced in parentheses in-text for reference to a video game title. Tables 4.17, 4.18, and 4.19 show the use of in-text citations and integration of source language in-text, according to the total essay score and component scores for material and correctness.

Table 4.17

In-Text Citation and Integration of Source Language In-Text (Total Essay Score)

Group_Total	**No In-Text Copying**	**No In-Text Correct Use**	**In-Text Copying**	**In-Text Correct Use**	**Total**
1 (High) n=17	4	3	3	15	25
2 (Mid) n=17	13	0	5	10	28
3 (Low) n=16	6	4	4	9	23
N=50	23	7	12	34	76

Table 4.18

In-Text Citation and Integration of Source Language In-Text (Material Component Score)

Group_M	No In-Text Copying	No In-Text Correct Use	In-Text Copying	In-Text Correct Use	Total
1 (High) n=17	3	2	3	15	23
2 (Mid) n=16	9	1	5	13	28
3 (Low) n=17	11	4	4	6	25
N=50	23	7	12	34	76

Table 4.19

In-Text Citation and Integration of Source Language In-Text (Correctness Component Score)

Group_C	No In-Text Copying	No In-Text Correct Use	In-Text Copying	In-Text Correct Use	Total
1 (High) n=14	6 (9)	1 (1.5)	2 (3)	11 (16.5)	20 (35)
2 (Mid) n=22	9	6	8	13	36
3 (Low) n=14	8 (11.5)	0 (0)	2 (3)	10 (15)	20 (35)
N=50	23	7	12	34	76

Note. Numbers in parentheses have been added to allow comparisons across level groups, as the groups are unequal in size.

It seems that the material component score reflected the most the test takers' integration of source language in-text and attribution to sources in-text because in Table 4.18, essays that used source language correctly tended to be rewarded (higher numbers in the upper right-hand corner and lower numbers in the lower right-hand corner), while most essays that included copied phrases or sentences with no in-text attribution to the source were downgraded in the material sub-score (higher numbers in the lower left-hand corner and lower numbers in the upper left-hand corner). This trend was also somewhat apparent in the total essay scores in Table 4.17 but not as clearly as in the material component scores in Table 4.18. Lastly, the correctness component scores in Table 4.19 did not reflect use of in-text citations and integration of source

language in-text because essays that did these two things correctly were not necessarily given higher correctness component scores.

Overall, the students' test-taking processes reflected test-taking behaviors that are expected in a construct of web-researching-to-write. The students' essays also displayed use of web sources through in-text citations, end-of-text lists of references, and integration of source text language. This was captured by the rubric and reflected in the test scores. However, the scores did not reflect the selection of credible sources to be used in the essays, which points to a need to further revise the rating rubric and/or rater training materials.

Comparison of Group Differences

Research question 4.4 asked how test performance is related to test takers' English learning, writing experience, and internet use. This was accomplished through an examination of correlations between test performance and (a) length of English learning, (b) source-based writing experience, and (c) internet use through quantitative analysis of post-test questionnaire responses (n=40). The correlation coefficients were all close to 0 and not significant (length of English learning: Pearson's r=-0.072; source-based writing experience: Spearman's rho=0.092; amount of internet use: Spearman's rho=-0.071). The mean test scores for extreme lows (n=8, years: 2-5) and extreme highs (n=11, years: 10-15) in terms of length of English learning were also compared, but the mean difference was not significant (t=0.207, df=17, p=0.838). The distribution of test takers' length of English learning is displayed in Figure 5.

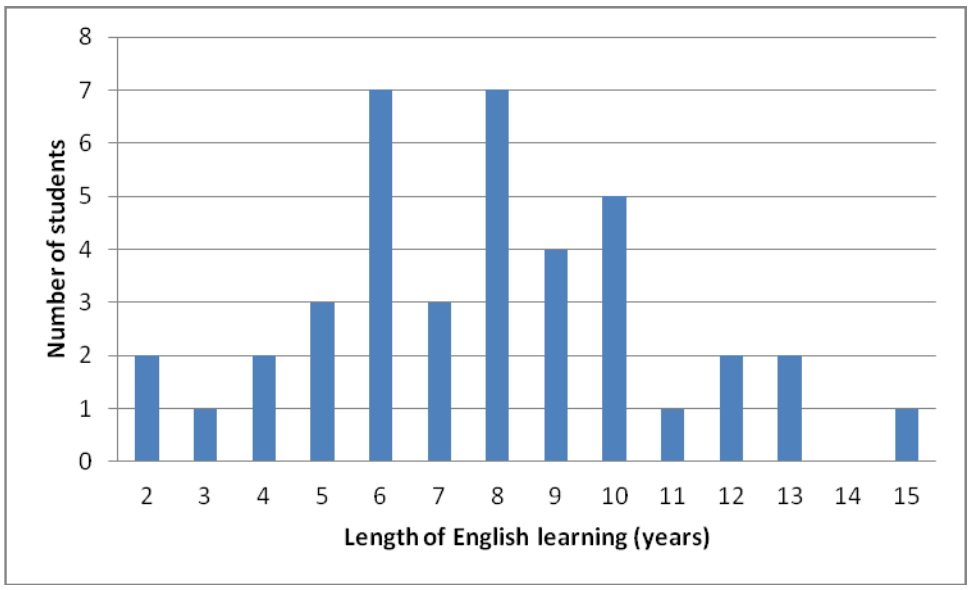

Figure 5. Distribution of test takers' length of English learning (n=40).

To conclude, the test takers' length of English learning, source-based writing experience, and internet use did not have a significant relationship with their test scores. It is possible that the test elicits performance that requires skills that were directly taught in English 101C and is thus not greatly affected by students' previous experiences with English learning, source-based writing, and the internet. The non-significant results may also have been due to issues with the veridicality of the post-test questionnaire responses. Having test takers complete the background portion of the questionnaire before the test may prompt test takers to respond more truthfully to the questions.

Extrapolation Inference

The extrapolation inference is warranted if the construct of web-source-based academic writing as assessed by the integrated writing test accounts for the quality of web-source-based academic writing performance in college courses, particularly with regard to those skills taught in English 101C. Backing was sought by collecting criterion-related evidence through (a) an

examination of relationships between students' test performance and self-assessment of their own source-based writing ability and (b) an examination of relationships between students' test performance and performance in a post-English 101C composition course that requires the completion of source-based writing assignments.

Criterion-Related Evidence

Research question 5.1 investigated how test performance is related to students' self-assessment of their source-based academic writing ability and students' performance in a post-English 101C course which requires the completion of source-based writing assignments. This was done through an examination of correlations between test performance (total essay score) and (a) students' self-assessment of source-based academic writing ability and (b) students' performance in a post-English 101C composition course (English 150).

Firstly, the relationship between test performance and students' self-reported confidence about citing sources in writing was investigated by computing the correlation between the students' total essay scores and students' responses to the post-test questionnaire item which asked respondents to rate their agreement with the statement, "I am confident about citing outside sources in my essays or research papers." The Spearman's rho was 0.450 ($p=0.002$, $n=40$), which was significant at $\alpha=0.01$ level (one-tailed) (see scatterplot in Figure 6 and box-and-whiskers plot in Figure 7).

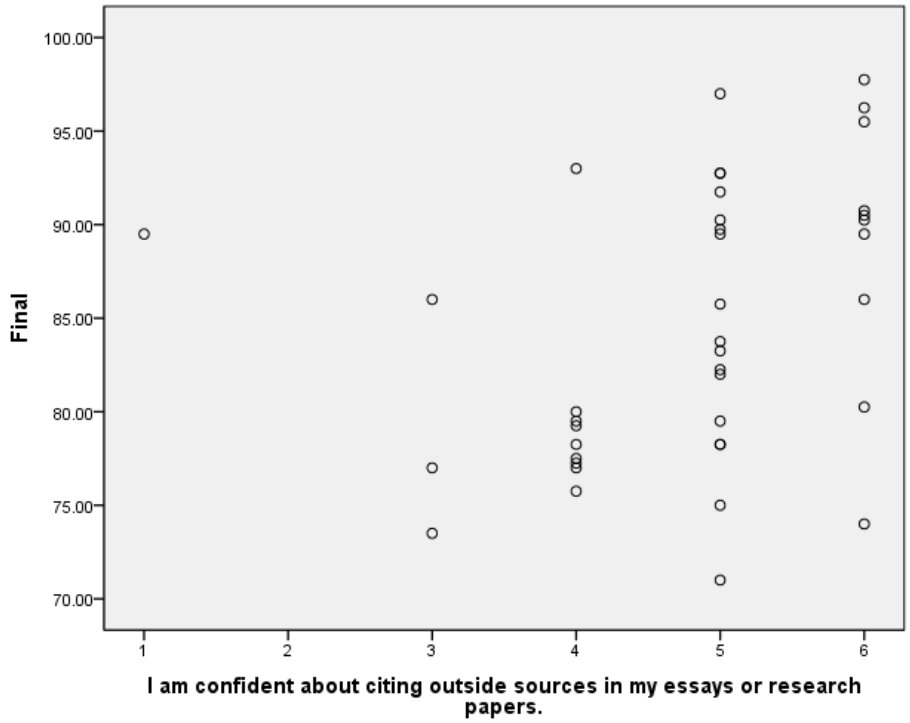

Figure 6. Scatterplot of final essay score and self-reported confidence of citing sources (n=40).

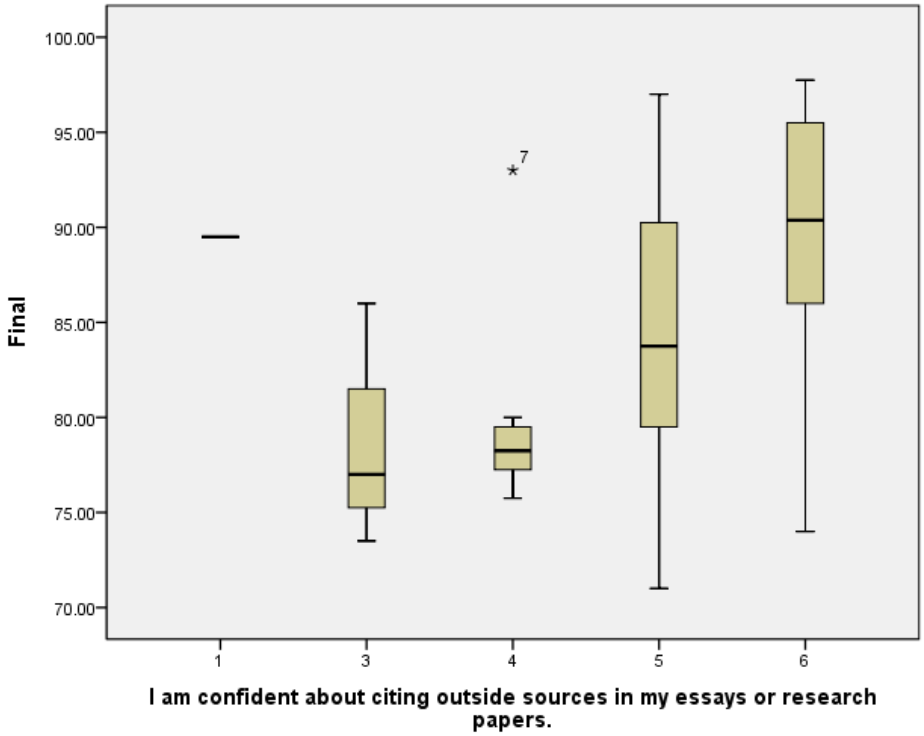

Figure 7. Box-and-whiskers plot of final essay score and self-reported confidence of citing sources (n=40).

Secondly, the relationship between test performance and performance in a post-English 101C course was investigated by computing the correlation between the students' final essay scores and the final course grades in English 150 reported by nine students who responded to the follow-up questionnaire. The Spearman's rho was -0.550 ($p=0.125$, $n=9$) and not significant at $\alpha=0.05$ level. A possible explanation for this finding is that the final course grade in English 150 is determined on the basis of not only writing assignment grades but also grades received for a visual assignment, an oral presentation, and a final portfolio as well as attendance. Therefore, the English 150 course grade does not reflect only writing ability. Another explanation could be that the nine students took different sections of English 150 because English 150 varies across section types in several different ways, including student composition, materials and assignments, instructors, and how student essays are evaluated. English 150 has two different types of sections at Iowa State University: regular sections which consist of mostly native-speaker American students and cross-cultural sections which consist of roughly half American students and half international students. In addition, there is the possibility of taking English 150 through the Des Moines Area Community College (DMACC). The break-up of the nine students according to type of English 150 taken is shown in Table 4.20.

Table 4.20

Types of English 150 Taken by Follow-Up Interview Participants (N=9)

Type of Course	**Student (Semester Taken, Curriculum)**	**Final Essay Score from English 101C**	**Course Grade for English 150**
ISU regular English 150	6 (S12 WOVE)	89.5	B
	22 (S12 place-based)	95.5	B-
	30 (S12)	83.25	B
	39 (F12 place-based)	97	B-
	42 (F12 place-based, but with more than half international students)	82	A-
ISU cross-cultural	45 (F12 place-based)	86.25	B+

English 150	48 (F12 place-based)	94.5	A-
DMACC English 150	3 (Su12)	89.75	B+
	47 (Su13)	86	A-

Follow-up interviews with nine students provide some qualitative evidence that what is tested on the final exam extrapolates to what students need to do later in English 150 and other courses. Firstly, all nine students acknowledged that source use is required in English 150 regardless of course type. Further, four students (Students 6, 22, 30, and 45) confirmed that source use is required in English 250. Source use is also often required in content courses, as confirmed by five students (Students 3, 22, 30, 47, and 48). Table 4.21 summarizes the source-based assignments in post-English 101C courses, together with representative comments from the follow-up interviews.

Table 4.21

Source-Based Writing Assignments in Post-English 101C Courses

Course	**Source-Based Assignments**	**Representative Follow-Up Interview Comments**
Regular/ cross-cultural English 150	Research essay	Well, I think yeah yeah, the last one yes, but it's more to, mm my instructor allowed us to search on Google, which is um doesn't require any like, she doesn't limit us to uh you only can find the sources from like scholarly articles. But she didn't limit us. She just let us use the internet however we want to. But so that's why I guess English 150 I did was ok. (Student 6)
	Expository essay	Most assignments is about something about our school or let's see yeah most of them is about the place or buildings or history about our school. So I should look up for most information online. (Student 39)
		Yeah, describing one of the building in campus, like from physical looking. And especially for this assignment, I did a lot of research job, search something from history and architecture style, something like this. (Student 42)
	Brochure	And we also and uh like I mentioned before, the brochure assignment. I do a brochure to, uh like to, about to introduce the

		Live Green Initiative. It's uh on-campus….So I also need to find some information… about Live Green. What's Live Green? And then who (believe? will be in?) Live Green? So yeah that's I guess it's uh such kind of source-based. R: Yeah, did you have to look for information? S: Yeah, a lot of information about who found Live Green and what Live Green do, what's its mission statement? And yeah. What's the regular activity they do every week, every month? (Student 22)
DMACC English 150	Comparison essay	S: I think the one essay I write is about the global economy, and I you know because I took the Economy 101 before, I used some source from that book and also researched some information of the website, and yeah. That's it. R: So what course was that for? The one about global economy. S: Yes, it was 150. The topic is global economy. It was about the comparison. (Student 3)
	Cause and effect essay	For the third assignment, I use some source from internet. And I just I Google. Like cause I can make assumption like what cause rude. I guess it may (in) those children learn from their parent and their education and something else I can, I then type the, do education related to the impolite? And search the related information. And I will use cite to, cite for it. (Student 47)
English 250	Research paper	S: When I take English 250, we're required to I was required to write a research paper, like uh I choose the topic it's that uh about the international economics. Why international trade is so important for the development of economics for uh in nowadays for the (developed) countries and (undeveloped?) countries. So it's uh it's hard because I barely know the knowledge about international trade, so I do a lot of reading and read some papers about this stuff and write the write a research paper, so. Yeah. R: Did you have to read articles? S: Yeah, read articles and I had to find the articles myself. I searched on Google, um Google, yeah. (Student 22) S: Yeah, at that semester. So I went to the E-Library on ISU and searched the paper I need, and in 250, instructor provided several database for us, so it will be much easier for me to find what I want, especially some database like for news, specific, specially. R: For news? Was it like Lexis-Nexis or something like that? S: Yeah. And some academic library or something. R: So you had to use the database and E-Library. S: And E-Library. R: Then what did you have to write based on those sources? Did you have an essay or a research paper? S: Normally an essay. (Student 30)

Content courses	Researching company information for Marketing or Accounting project	I take Marketing 340 last semester, and we do a final project….It's uh we have small groups consists of five to six people and we each choose a topic like, oh yeah, we do project like to compare two food company, Dunkin Donuts and Krispy Kreme. And we need to search information. Uh actually one is successful company, one is unsuccessful company, so we compare two company. So we search information, the background of two company, and compare their performance, sales performance, or such things, yeah. So I do a lot do a lot of I mean searching for information. (Student 22)
	Researching for humanities course essay	S: Yeah, and I also used cite for the Religion 205's project. Yeah because, The Matrix, because you should find some other one's idea of religion and Matrix, so I find then and use (?). It support your idea just. R: Yeah. That would be a really good assignment to use sources. S: Yeah, you should use source, from book, from the like textbook, the book from library, and internet, journal, yeah, you should use some idea from other people to support your idea. (Student 47)

Secondly, eight out of the nine students found at least some aspects of what they learned in English 101C to be useful while completing writing assignments in English 150 and other courses. Aspects of English 101C that students found useful included essay structure, grammar, referencing, and writing practice (see Table 4.22 for representative follow-up interview comments). These aspects were assessed in the final exam for English 101C.

Table 4.22

Usefulness of English 101C Content for Future Writing

Aspects of English 101C	**Number of Students**	**Representative Follow-Up Interview Comments**
Essay structure	4 (Students 3, 30, 42, 47)	Yeah, it's a long time, but I think the 101C is a more basic writing class. And we learned some like the structures of the essay and those stuff is more like the very basic one and not like the academic one like the 150, not like that. And I think it is pretty useful. The way we write an essay and write a report is easier for me to you know to build up structure of the essay and follow structure. And yeah I think it's pretty useful. (Student 3)

		S: It's quite helpful for the basic structure and I think the basic form of my essay, what should it looks like. R: The basic essay structure? S: So at least I know what I should exactly write, what I should not write. And I also use that in other courses when I write some other assignments. (Student 30)
Grammar	4 (Students 39, 42, 47, 48)	So the grammar I learned in 101C is useful… (Student 39) I learned a lot from 101C like how to check the grammar, use the citation, and how to write a paper like structure it, configure it, so I use it to the pa, uh to other writing, so when I first write, the first draft, I will like I will check the grammar sentence by sentence, and then I will consider if add more detail or delete some details. (Student 47) Another thing is the grammar error because I got bigger improved in my grammar when after I take English 101C… (Student 48)
Vocabulary	1	The vocabulary is useful. (Student 39)
Source evaluation	1	S: Google to find some information related and then choose if it can be work. Mm hmm. Cause some information is not work. R: Right. S: And also I will take care if it's like it's a journal or it's a paper or it's some just like blogs. It's important. So it's if they're reliable. Mm hmm. R: So you evaluated the source? S: Uh huh. Cause you told us the org, edu are reliable. R: Yeah, they are usually reliable. S: And like wiki encyclopedia. R: Wikipedia? S: Wikipedia is not that reliable as other. (Student 47)
Referencing	3 (Students 45, 47, 48)	I tried to apply what I remembered, and I think I did well. When I was told to do things like MLA, I just exactly followed them. (Student 45) For Accounting 254, because I need to do lots of research and at the end, I need to put lots of reference that I have researched. So this is one thing. (Student 48)
Citation-producing website	2 (Students 47, 48)	S: Um, I don't think so. Oh yeah, they talked about it. But I forgot. Cause you can like use Citation Machine and you have like different style so if ask you this style, you can circle this style. R: Oh so the instructor told you about Citation Machine? S: Mm mm, no, you told us. R: I did? Yeah, I did!

		S: Yeah. R: Citation Machine and KnightCite. S: Yes, I think the 150 asked is MLA. (Student 47) R: Oh the Citation Machine? Are you still using them? S: Yeah, yeah, yeah, yeah. I use it in my last in my computer science course, Accounting, Theater 110, and other. (Student 48)
Becoming comfortable with writing through practice	2 (Students 45, 48)	We did a lot of writing practice, so I was more used to writing courses by the time I went to the 150 class. Yes, I got a bit more used to writing. (Student 45) S: Before I came to US, I was struggle for the TOEFL exam. So I prepared the writing a long time. But every time I start to writing the TOEFL, TOEFL writing, I was nervous and don't know how to do. Sometimes I will, my brain is blank. I don't know what's I need to use the example. But after the English 101C, in the final exam, I feel very peaceful. Just writing. So I think I got really big improvement. R: You felt more confident? S: Mm hmm. So after this class, I also got big improvement for our future writing English writing. (Student 48)

One student (Student 6), however, did not find the content of English 101C helpful for completing writing assignments in later courses. The reason was attributed to the gap that exists between English 101C and English 150, the latter course being a regular section for Student 6:

> I think it [content of English 101C] wasn't that helpful at all because…yeah, it's completely different. I was really shocked when I received my first graded assignment on English 150. She told me that she liked my essay very much. She say it was very narrative, it's like story-like. And I end up getting a B-, so I was like oh I was quite shocked when I saw the grade, so I know that the gap between English 150 and English 101C is very wide. It's definitely a different thing. (Student 6)

With the exception of Student 42 whose section had more international students than American students, the other four of the five students who took a regular section of English 150 pointed out the gap that exists between the expectations in English 101C and English 150. Similar sentiments to Student 6 were had by Students 22, 30, and 39, who also took regular

English 150. For example, Student 30 described how he had a challenging time making the transition and initially adjusting to English 150:

> I had some problems at the beginning of English 150. I just got a C at the first two assignments, but after that, I reviewed the comments on them and I make some changes and developments, so the rest of my assignments were pretty good. And when I went to English 250, I did quite good in that class also. (Student 30)

Student 39 was enrolled in an English 150 section where he was the only international student. He did not find all aspects of English 101C to be helpful while taking English 150 because he perceived the assignments in the two courses to be different:

> S: Because I think this like English 101 is quite different from English 150. I think it's a totally different level, so they write different kind of essays. So the grammar I learned in 101C is useful, the vocabulary is useful, but the strategy of writing the essay in 101 is, it's not useless but it's not really helpful for English….I studied the TOEFL, so the assignment in English 101C is quite similar to the TOEFL writing and test, but the English 150 like kind of like it's not free style but it's not what I'm good at, like that.
> R: So you had new assignments. New types.
> S: Yeah, new sort of writing tasks….Yeah, it's yeah it's different kind of writing requirements….The strategy in English 150 is different because I can tell the requirement is, the topic requirement is quite different from 101. The most different thing is the strategy, how to structure my essay, like that….Before I took 150, I used to write something like give a, um like the introduction of the topic and some example or some personal experience or something to support your topic and the last paragraph is to conclude what you said or rewrite all things you have mentioned in (previous paragraphs), but this strategy cannot use in English 150 because it's not to to like how to say whether you support or not support about something. It's like ask you to describe something, so this strategy cannot use in this situation. You need to I don't know I just not want to like it's more like to be more specific or to write the detail of something, which is I I didn't practice in the previous….Some of the assignments in English 150 is like to describe something or to introduce the history or the culture of something. And because the information is from the internet so you need to paraphrase something or conclude the main idea of it, so it's it's harder. (Student 39)

He further explained how the requirements are stricter in English 150:

> I think English 150 has more strict requirement for the citation….I still remember that I like Assignment 3 or Assignment 4 in English 150 need us to cite the source we get from the internet, and I did that, but it's in wrong format, I mean, so it's like in the rubric that the citation that way, I got zero in that, so yeah. (Student 39)

Thirdly, seven out of the nine students believed that source use should definitely continue to be taught in English 101C. One of the main reasons for the positive responses was that many international students are not aware of the necessity of proper use of sources in writing when they arrive at a U.S. university:

> I don't know that maybe in America most students learn this stuff in high school and yeah or middle school, but I can just say in China we didn't do this stuff, so actually I don't know we need, I know some journal have the citation, but I don't know it's necessary when we write the essay... (Student 39)

> Mm, yes, because that is what you do when you first start English, so you need to know it to be able to keep citing later in future English classes. I think it's difficult to get used to citing because we don't do it much in Korea. (Student 45)

> Yes, because lots of international students, they take this class, but they don't, some of them, they didn't know that in their original country. So it's very important for international students to know how to use this and citing sources. Yeah, it's very important to write a paper in United States. Yeah, and college life….For international student, it's tell them this is very important part....I think this things need to be emphasized to the international student. It's very helpful for the in future writing. Yeah. Now my writing is better and got better grade….I think it's very practical thing and practical skill. (Student 48)

A second reason was that repetition is helpful:

> Yeah, I think so. Both, it's ok to teach it in both 101 or 150 because something those stuff is very easier to ignore sometimes, and you use some source from the website and you forgot to cite it. It always happen. So I think that if you teach it in both course, maybe it will remind us a little bit. And yeah, I think you can do that in both course….Yes, sometimes it happens also in my other course, like my majors, sometimes they learn the same stuff, like we learned from physics or chemistry. So it happens. So yeah, I think it's ok for both course to teach those stuff. (Student 3)

Another reason was that it would give students a head start in preparation for the extensive use of sources required in English 150 and 250.

> I think I think um yes, it should be taught in 101C because from English 150 and 101C, it's it's totally different thing, the difficulties are like from easy to super difficult. So it's yeah so I think using and citing sources should be taught so that

> uh to give the students a head start of what's gonna come when you take English 150. So yeah. (Student 6)

Two students, on the other hand, did not see source use as a necessary aspect for English 101C. This was because source use is taught again and again in future English courses. These two students thought that an introduction to the concept is enough:

> I'm not sure because in most of my classes, they will talk about how to use and cite sources. In 150, 250, and now I'm taking 302….Yeah. So I'm not sure if it's really necessary to do that. But at least you can mention that when you just cite the reference. I remember in a presentation in English 150, and it's a presentation and of course we had to use some pictures online, and we need to hand in a report about our presentation, and most students, they didn't have a Works Cited page, but we did. So we get the full score in that presentation…so maybe in 101C, you (will/would) just need to mention. (Student 30)

> Not really….Just because in English 101C, it's to teach the students like how can you write a sentence completely, I mean basically from the, and the grammar, organization, something like this. This is the most important idea in this class. For the cit-, source, mm I mean using source is, you can give them a few class to know about this topic, but not really more because in English 150, you will learn this. (Student 42)

To summarize, the relationship between test takers' performance as represented by total essay score and their self-reported confidence about citing sources in writing was shown to be significant. However, the relationship between test performance and performance in courses post-English 101C was not shown to be significant, although follow-up interviews with nine students provide some qualitative evidence that what is tested extrapolates to what students need to do later in English 150 and other courses. It would be worth conducting the latter comparative analysis again using grades received on specific source-based writing assignments in post-English 101C courses instead of overall course grades.

Utilization Inference

The utilization inference is warranted if the target scores obtained from the integrated writing test are useful for making decisions about final exam grades and appropriate curricula for students in English 101C. Backing was sought through an examination of (a) students' perspectives on whether they had equal opportunities to learn or acquire the ability to write from internet sources in English 101C and (b) students' and instructors' perspectives on the usefulness, clarity, and interpretability of the score descriptors.

Equal Opportunity to Learn

Research question 6.1 was "How equal did test takers perceive the instruction and preparation they received before the test?" The perspectives of test takers were obtained from the follow-up questionnaire and interviews. Students were asked the question "Do you think the students were given equal opportunity to prepare for the exam?" Except for Student 3 to whom I did not ask the interview question, the other eight follow-up interviewees responded that students had equal opportunity to learn and prepare before the test. In fact, the interviewees thought it quite obvious that everyone was given equal opportunity and that the responsibility was on the students who were absent or who did not pay attention in class. The following are three representative quotes from the follow-up interviews:

> Yeah, I think students are given equal opportunity to prepare for the exams. Yes, because everybody should have should have should be equal, right? (Student 6)

> S: Yeah, of course, yeah. We all have the same instruction and the same opportunities to learn to practice. Yeah.
> R: Ok.
> S: And you will illustrate everything in class and uh so that's pretty fair.
> R: I guess if you are in the class, we all had the same. If you were absent, then maybe not.

> S: Oh no no you will also send emails and something like that and that's ok. (Student 22)

> Of course. Because it's very it's very equal. Every student got opportunity to take this exam. I didn't notice any other cheat any I don't know. (Student 48)

Furthermore, test takers perceived the instruction in English 101C to have been adequate in preparing them to take the final essay exam. One of the items (Item 6a) on the follow-up student questionnaire asked students to rate on a scale of 1 to 6 how much they agreed with the following statement: "The instruction received in 101C was adequate for taking the final essay exam." The responses were positive (mean 5.444, standard deviation 0.882, range 4-6, mode 6, median 6). An open-ended question (Item 6b) asked the respondents for the reasons for their response to Item 6a. Except for Student 6 who gave no response, the other eight students provided mostly positive comments. The following are four representative written comments:

> The instruction gave us more idea about writing. It help us to build a writing structure before we write essay. Also, it teached us some writing skills. (Student 3)

> My writing skill got improved through every assignment before the final essay exam, and the practice in class including writing, reading and team work also consolidated my writing skill. (Student 22)

> The final essay exam is similar to the assignments. I can follow the feedbacks of the previous assignment to prepare the final essay (Student 39)

> What I have learned in English 101C was useful in the final essay exam and in the future English class. (Student 48)

The nine students were also asked to elaborate on their questionnaire responses during the follow-up interviews. The following are four representative quotes:

> I think it was on a scale from 1 to 6, I think it's, I would give 5 because it's important that we how do I say this, um it's important to have a final exam so and of course it's definitely useful because I think English 101C, the final exam summarize the entire course. So I think it makes you, you have to organize your thoughts and thinking on so many assignments that you did on the in the 101C course. You have to know by now what kind which is your strength and which is

> your weakness, so I think it's probably adequate for taking the final exam. (Student 6)
>
> Yeah. We have done many other assignments in 101C and you have graded I think somehow based on similar rubric like this, so I'll understand what I should do what I should focus in that final exam through all the classes. (Student 30)
>
> I think so….Cause I think we learned how to write a paper from 101C and we know how to organize it, and we can if we uh we should first have idea, like the thesis, and if we have thesis, we can give several topic sentences to support the thesis and we can add detail, example to make the paper vivid, so, so it's not too hard for you have experience of writing. (Student 47)
>
> S: During English 101C, I learned a lot about how to make reference and how to organize my idea well and how to make it make the form like two spaces and I don't know.
> R: Like indenting paragraphs? Double-space?
> S: Double-space and New Times Roman and font 12. So very useful to me because I didn't notice that um notice before taking this class. So it's very helpful to taking final essay exam. Very adequate. (Student 48)

Two test takers had also made very similar comments during post-test interviews regarding the adequateness of the instruction in English 101C for completing the final essay exam:

> I think it's a, it's a good final test for English 101 students to cover. And it will help students to uh review what they did learn in class. Some strategies about organization, some about the citing, also about how to search information, so I think it covers a lot. (Student 29, post-test interview)
>
> S: I think I feel pretty good with this test because I used the knowledge that I just learned before test to use it, and yeah I think I used it pretty good. Like I used the thing that you teach me in the class in the exam. Yeah, and I feel like my writing is better.
> R: Like comparing to the beginning of the semester?
> S: Yeah, it's a lot better cause before this I don't know what is the, I don't know the thesis statement. Or like I have to have introduction, conclusion cause I always write the same thing. Yeah, so like okay body body and then blah blah blah, no introduction and conclusion. (Student 37, post-test interview)

Usefulness, Clarity, and Interpretability of Score Descriptors

Research question 6.2 was "What did test takers and experts think about the usefulness, clarity, and interpretability of the score descriptors?" An analysis of the follow-up student questionnaire responses and interview data from test takers and instructors provided the answer to this question.

The nine test takers who participated in follow-up interviews were provided with a copy of the score descriptors to read and were asked to rate the usefulness, clarity, and interpretability. The questionnaire responses are summarized in Table 4.23.

Table 4.23

Follow-Up Questionnaire Items for Score Descriptors (N=9)

Item	Mean (on Scale 1-6)	Standard Deviation
5a. How useful are the score descriptors for understanding strengths and weaknesses in writing?	5.111	0.928
5b. How clear are the score descriptors?	5.444	1.130
5c. How easy to interpret are the score descriptors?	5.111	1.269

Note: 1 – strongly disagree; 6 – strongly agree

The follow-up interviews further elicited comments from the nine students regarding their perceptions of the usefulness, clarity, and interpretability of the score descriptors. First of all, regarding the usefulness of the score descriptors, six of the nine students provided positive comments about being able to tell which areas are strengths and which areas need improvement. The following are two representative comments:

> Yeah, if I received a table like this and I know which score I got, I think it's easier to see the you know which level of my writing because for example, maybe the organization maybe I got a 27, and some place maybe like the correctness, I got the a full marks and I will know that I need to work harder on the organization stuff. And because it's very clear, like the 27 include what type of the writing and I will know now how my writing looks like and how to improve it to a higher grade. (Student 3)

> The scores are useful because um in each box it's clearly stated uh well, so it makes me so like if you gave me on either one of the category, it makes me to know much more clearly about where my strength is and where my weakness is, so that's very useful….It makes me understand that how well I did on final exam. Of course that's definitely that's certain. So, yeah I would definitely use that information from this from the rubric that I've given to do to improve much better on the next the next or whatever English course I'm going to take in the future. I'll definitely use that as a reference. (Student 6)

Secondly, eight of the nine students also made positive comments about the clarity of the score descriptors, while thirdly, regarding the ease of interpretation of the score descriptors, two students made positive comments and five students did not make a comment during the follow-up interviews. The following are three representative student comments on clarity and ease of interpretation:

> S: So I (see/use) the rubric and (saw/start this) made in a great detail, so it's very illustrated so I have a clear uh clear source about what should I do to get a good score or so yeah. Yeah, so that's it's very long and then lot of like uh it's very clear yeah.
> R: There's a lot of detail?
> S: Yeah, a lot of detail, so after I read through the rubric, I think I don't need to ask any other questions concerning the grading criteria, so it's very clear. (Student 22)

> And for the part B, I choose it is very clear cause it write down all the detail, what score you can get, and what the requirement for each one. For the part C, I think it is very easy to understand the description cause it's not too hard. Yeah. It's write very clear all the details you should all the detail you should do. (Student 47)

> R: Are the descriptors clear to you when you read the boxes?
> S: Every time I receive this grade, I will see why I got this grade and depends on this things, so it's very clear to me to know which part I need to still need to improve. So I always got grammar error (knock off a grade?). It's very clear.
> R: And easy to interpret?
> S: Yeah, it's very easy to interpret. It's um everything is list here and what I need to do to improve. Yeah, so it's easier to see this (grade). (Student 48)

On the other hand, there were some negative comments and suggestions made about the usefulness, clarity, and interpretability of the score descriptors. Firstly, three out of the nine students had a negative comment about the usefulness of the score descriptors. Student 39

thought that the descriptors were too general, and thus, more personalized feedback needs to be given:

> The rubric it's like um how to say um it's like a general idea. It's not just for me like maybe I have some problem in some items of this grade. Yeah. Maybe I did good in the first one and second one but really bad at the last two, so I mean for example, so it's not personal, so I'm more like some feedback in my essays, so I can tell which part I can improve or which part I can revise in the assignment, but according to this [rubric], I just for me I if I before I write my essay, I can follow all this all this thing to write my essay, to make sure that I meet all this requirements, but after that, I don't think I can tell anything. I can just see the score of it. (Student 39)

Student 42 made a very similar comment about the specificity of the descriptors:

> R: So for example, like after taking the final exam, the instructor would give you the rubric with like certain boxes circled. Then would you be able to interpret your score?
> S: You mean, based on this I can know what score I can get?
> R: Yeah, and which areas you need to improve and which areas you did well in. Are you able to get that information, do you think?
> S: Basically yeah I think it's not enough. Like ok if I got uh so for example for the material part, if I got a scores on uh level 3, right? And all the things there's just a lot of the subtopic in here, you see? And it's all these subtopic are equal? Like which subtopic is more important in this section?...Yeah, check within box and show me show me what is the weakness part I cannot do really good, and before the final exam, I think for the rubric, it's kind of easy here and it's really good, it's really (?) to tell student what do you think which part you think will be the most important part when you grade my final exam….So I know that um when I write my final exam, what part which part I should pay more much more attention to. (Student 42)

Student 42 also doubted whether he would use the rubric after the final exam:

> Um basically to be honest, it's not, I will not use it again….Yeah, it's [the semester's] over, I, who actually use this…If it required us to use, it's basically like the other assignment. After you get feedback from the instructor, you need to check your rubric and which part you didn't do a good job and which part you still need to make more sense? Something like this. (Student 42)

Student 45 pointed out that there was too much text and detail in the descriptors, making it difficult to address everything:

> Since it [the rubric] says what is lacking, I think I would put my effort into correcting that….There is a lot of content in one box, so I can know well what I lack, but it is difficult to correct everything following all of the content….I think there is too much text…So if I just read this and I read about what I lack based on the ratings, but there is too much content, so it's hard to remember what I lack? Because there is too much…But I think it is good to know what is lacking and what is good. (Student 45)

Secondly, regarding the clarity of the score descriptors, one student, Student 30, pointed out that some terms, particularly quantifiers, need to be clarified: "I give a five here just because when I saw here, it's little and some, sometimes it's may easy to confuse. I don't know the actually the actual you know standards for what is little, what is some."

Thirdly, regarding the ease of interpretation of the score descriptors, Student 6 felt that she needed elaboration of the score descriptors in general:

> To interpret the score descriptors, if whenever I read them myself, I think it's I will need to find to spend some time to try to get an appointment to talk to my instructor and ask him or her to elaborate what um how the score descriptors, elaborate the score descriptors to me. So yeah that's why I think it's it's more like on the neutral side. (Student 6)

Similar thoughts were expressed by Student 42: "As for me it's not easy to interpret like this things. I'm not a good writer so if I if some people ask me to interpret his rubric to him, I just read these things again."

The qualitative responses in the follow-up test-taker interviews can be summarized and quantified as follows: usefulness of score descriptors (6 positive; 3 negative); clarity of score descriptors (8 positive; 1 negative); and interpretability of score descriptors (2 positive; 5 no comment; 2 negative). The follow-up test-taker interviews further revealed a need to provide the rating rubric to the students in advance before the test and to explain and discuss the score descriptors to ensure that all students understand what is expected at each score level for each

criterion. It may also be necessary to clarify some terms in the score descriptors, for example, the quantification of "little" and "some" as pointed out by Student 30. An additional procedure in the rating process that can make the score report more useful for the test takers is to highlight or check specific bulleted points within the boxes in the rubric. This could clarify to the test takers why their essay received a certain level for each criterion and what specific aspects need improvement, thus personalizing the feedback provided by the marked rubric.

The five previous and current English 101C instructors were asked the interview question "How useful, clear, and interpretable do you find the score descriptors?" Overall, the instructors found the score descriptors in the rubric/score report to be clear particularly for the instructors, but many suggestions were provided for revision of specific details of the rubric, most of which were discussed under the evaluation inference in the systematic rubric development section.

Instructor 2 thought that some terms in the score descriptors should be specified, while Instructor 4 thought that certain aspects in the rubric should be moved across levels:

> Yeah. I think fine….I think that's fine. I mean that's clear. But again like I said, for example, length must be specified, you know thesis must be specified, cohesion must be, transition must be specified. Because there are, in this genre, you are targeting certain transition words. They are not everything, so I would specify them. (Instructor 2)

> Yeah if you're talking from an instructor's perspective, this is clear because an instructor will understand most of the concepts…the things are clear now, but the only thing that will be misleading is something that is satisfactory but it's not actually satisfactory. So if you make those changes, it will fall under the right headings and that will be fine. I think they [students] will understand. (Instructor 4)

Instructor 5 suggested a need for two separate versions of the score descriptors—an instructor version and a student version—with the student version using simplified language:

> Yeah as a teacher, I would have no problem to use this rubric and to understand and interpret the information that's given in all the descriptors. But from the student's point of view, for some parts, they may need some additional

> explanations to fully understand what the descriptors mean. And those are for example, like some technical terms that are widely used in our field applied linguistics but that probably wouldn't be necessarily trans or clear to students, so like sentence variety or sentence complexity, um then could vaguely know what this means, but maybe not clearly. Those kind of things may need to be further explained for students….for some words may not be um easy for for students at this proficiency level, so like covert overt or like impede or those words may be a little hard for especially lower proficiency level students I think. Or obscured. I mean they could refer to the dictionary if they don't really understand what those words mean, but if there is a way to simplify these words, then it will be more preferable for students I think. For teachers, no problem. (Instructor 5)

Particularly for the student version, Instructor 1 suggested the use of examples to illustrate some of the descriptors:

> R: Do you think the students will be able to understand the descriptors?
> I: I think so, but always um you can use some examples to students. Yeah. (Instructor 1)

Instructor 3 thought the score descriptors were useful and interpretable, but suggested that there remain issues with clarity because of the potentially unclear boundaries between levels.

> It's it's certainly it's useful. It's useful because it's very detailed. And it it's very straight-forward. Clear, interpretable, I think it's interpretable because the wording is easy. The language is not hard so I think it's very easy to understand for both the instructors and the students. But clear, well maybe the problem again is about boundaries. So yeah, that's an issue because if you give students a 26 in this column, they will absolutely think their paper is posited here, but sometimes it's not there. It's misleading. Some of the things they do not need to focus on because they are actually good, but it's here, it's not good. So I think that's an issue. (Instructor 3)

To resolve the boundary issue, Instructor 3 suggested highlighting individual elements within the score descriptors across all levels and deciding on a score that is within the range after doing "some calculation in the end" without being restricted to choosing one level per criterion. To further enhance the interpretability and clarity of the score report, Instructor 3 recommended using complete sentences and replacing big words with simpler words. She also suggested the possibility of developing software that allows dragging and dropping of elements of the rating

rubric's score descriptors into a comment box so that students would see a paragraph written by the instructor. She believed that this form of feedback would be more interpretable for the students. Lastly, Instructor 3 thought that instructors could discuss the rubric and score descriptors with the students before the exam:

> Maybe one thing that, maybe one thing that we can do as instructors is we can go through the rubric together using one class period before the final exam. And then you know go through the rubric and you know explain the hard vocab and have them take notes. Have them study the rubric hard before they actually move on to the final exam in order to make sure to guarantee the clarity of clarity of this rubric and the interpretability. (Instructor 3)

In summary, there was a large overlap in the students' and instructors' comments on the usefulness, clarity, and ease of interpretation of the score descriptors. More than half of the nine students and most of the five instructors thought that overall, the descriptors were useful, clear, and easy to interpret, but there were several improvements that could be made to the score descriptors themselves as well as in the procedure for implementing the test and marking the rating rubric.

Implication Inference

The implication inference is warranted if the consequences of using the integrated writing test and the decisions that are made are beneficial to the stakeholders, who are the students, the English 101C instructors, and other instructors who will teach the students in the following semesters. Backing was sought through (a) a washback study to investigate whether the test will have a positive influence on how academic writing is learned and taught and (b) the record of time taken to rate essays during rating sessions to ensure that score reports can be distributed to students in a timely manner in order to further promote positive impact of test use.

Washback Effects

Research question 7.1 asked what the washback effects of test use are on instruction and learning. The data that were collected and analyzed to answer this question were the follow-up student questionnaire and follow-up interviews with nine students and interviews with five previous and current English 101C instructors.

An open-ended item on the follow-up questionnaire with test takers (Item 7) asked the students, "When you learned about the requirements and details of the final essay exam in English 101C, how did you prepare for the test?" The nine respondents' responses are summarized in Table 4.24.

Table 4.24

Summary of Test Takers' Preparation Activities before the Test

Type of Preparation for Test	**Number of Students**	**Student Responses on Follow-Up Questionnaire**
Review feedback on prior assignments	3	I…read through my prior assignment and comments on it. (Student 22) Review all previous assignments. (Student 39) I spent some time on all papers that I wrote before final exam in English 101C. (Student 42)
Think about organization or structure	3	I did a lot of thinking by how to get my thoughts into writing. I had some difficulty trying to organize my thoughts since organizations are one of my weaknesses. (Student 6) I outlined the structure in my mind before the final. (Student 22) Prepare the main idea and structure of my essay. (Student 30)
Read references	1	Read reference which can contribute to my exam. (Student 30)
Prepare vocabulary	1	Nothing or prepare some vocabularies which would be used for the final essay exam. (Student 48)

No preparation	3	No response (Student 3) I didn't really prepare because the essay question wasn't given so that I decided to write anything relate to the question when the exam starts. (Student 45) I didn't prepare test a lot because I think writing doesn't need prepare a lot. (Student 47)

During follow-up test-taker interviews, students were asked to elaborate on their survey responses. First of all, three students commented that they reviewed the instructor's feedback on prior writing assignments in preparation for the final essay exam. For example, Student 22 explained that he reminded himself of the mistakes to avoid during the final exam by reading the instructor's comments on his essay assignments:

> Ok, I, first I read through my assignment before the final, like the four assignment and the comments and I, I'm sometimes, when I just finish an assignment and receive my score and read the read your comments I will know oh this is my weakness and then I'll change it, but I mean after few weeks, I may forgot these things and I'll write another assignment, I will made the same mistakes, so before the final exam, I mean there are all these weakness in my mind and do not let them happen in the final exam, and I will try to yeah that's all about that I guess. (Student 22)

A second major type of preparation activity that three students brought up was thinking about the organization or structure of the essay:

> Yeah, so when I learned about the requirements and details, of course I did a lot thinking when I was back when I went back to my dorm, and I've try organizing my thoughts and I also set up a mind map, so as to be more well prepared for when I attend the final exam. So that I could know what to write and how to write and how many paragraphs I should write in that essay. So yeah that's probably how I prepared for the test on that day. (Student 6)

Lastly, three students responded that they did not particularly prepare for the test. Student 3, for example, gave the following explanation for why he did not prepare for the test:

> What I know is we have to write an essay in two hours, in one or two hours, and you told us it is, you will give us the topic and the essay, um, I actually didn't do a lot to review those stuff because it's a writing skill. It's not like math or something I need to practice. (Student 3)

Based on the students' responses to the follow-up questionnaire and the comments elicited during the follow-up interviews, the use of the integrated writing test seems to have had a positive washback effect on learning. Two thirds of the nine students prepared in some way for the test, and the preparatory activities were meaningful in terms of the requirements and expectations of the test.

The five previous and current English 101C instructors were the other group of participants that were interviewed regarding washback effects of the test. The instructors were asked two questions. First, they were asked how they would feel about using the test as a final exam in English 101C. Then they were asked "if you were to use this test as a final exam in your class, what might be some potential washback effects of using this test on your teaching and test preparation?" Since the test was only used in English 101C sections taught by the researcher, the instructors speculated about what the possible washback effects might be if they were to use the test in their section of English 101C as a final exam.

In response to the first interview question, four of the five instructors said they would be willing to adopt the test in their English 101C courses. Instructor 2 expressed the most positive response: "Yes, yes, yes, three yes's…You still have them [students] in test situation, but at the same time, you give them enough freedom….I think this is very authentic." Instructor 1 said, "I think it's a very good option," while Instructor 5 saw the positive in the possibilities of topics that can be chosen for the test:

> I think this is a really great test because…if we allow them [students] to use additional resources from the web, then the scope of topic that we can choose for

> that is wider. So we could use some more engaging topics which will motivate students to actually write. (Instructor 5)

Instructor 3 was a little apprehensive about the difficulty level of the test, but was still interested in adopting the test:

> I think it's going to be challenging. Yeah, but then on the other hand, I think it's it's focusing on one very critical skill, um about English writing, so I think it's very, if I were asked, you know if I were you know informed of this option, I will try it. I will try it. Um almost immediately. (Instructor 3)

The one instructor who was against adopting the test was Instructor 4. He preferred to give a take-home final exam for which students revise an essay and write a reflection on what changes had been made and why. This is because a timed essay exam does not give students enough time for revision and reflection, which are two key aspects of writing for Instructor 4:

> I wouldn't use this as a final exam…as I said, my thing would be to give them an exam that is multimodal….I will make it a take home. I'll make it a take home in the sense that writing for me is also revision….And the timed essay does not really give them time for revision….So I would rather give them a take home essay as a final essay. And then look for all these things that I'm looking for. In that case they have room. They have time to really demonstrate what they have learned in the semester…so that's why I wouldn't do a timed essay as a final essay. (Instructor 5)

In response to the second interview question, the five instructors foresaw several potential washback effects of using the test. The most common response, given by three instructors, was that they would introduce the concept of documenting and using sources near the beginning of the semester rather than towards the end. For example, Instructor 1 said he would introduce both the concept of searching for sources and the rubric that incorporates the concept of source use at the beginning of the semester:

> I: I think there would be some good positive washback effects because I would, at the very beginning of the semester, I would let students know, in this semester we're going to use this method, this approach in writing, so you have to use, you have to go to search some stuff yourself. I think that this is good. Yeah so when the students are used to such a writing approach, yeah. That's a good effect.

> R: Do you think you would prepare your students any differently?
> I: I do not really have to if I use the rubric at the very beginning to let students know why you, how I evaluate your paper. So that's kind of because that's the goal of the teaching, to help students master the skills of using internet resources, right? So yeah. The teaching and the evaluation process is actually a kind of preparation process for students. (Instructor 1)

Similarly, Instructor 2 said he would introduce the final exam at the beginning of the semester:

> If you tell them [students] this is the final exam right at the beginning of the semester right? And you draw the line of the objectives at the beginning of the semester and they know what they can do or they have to do at the end, so they will study paper 4 more carefully….I think it will have a good washback effect on both instructors and also students, especially instructors because it will push them to do something towards this goal because you want your students to be able to do this final. If everybody fails this final then it means you didn't teach anything. So that's the, that's a good washback I think. Push. I just mentioned it. I would be very happy, and I'll feel very happy for other instructors too because it is something they need to push for. It is difficult but they shouldn't give up. (Instructor 2)

Instructor 3 also said she would introduce documentation of sources near the beginning of the semester rather than in week 15, which is when she currently introduces the skill:

> If we're using this test and I think I would probably want to revise the syllabus a little bit. Maybe move things around. Maybe I will talk about documentations right away in Unit 1 or Unit 2….I will not wait till you know the last two weeks of the semester and talk about a very new thing and then in the final exam, I will te, you know will test this skill. Because if it's covered in the final exam, I assume I would you know unconsciously talk a lot about this skill because I would afraid that I would be afraid that they um go to their final exam with a lot confusions in their mind. So I would probably you know the focus will shift toward you know the documentation skills more, much more….I probably will teach it at the beginning so they would get more chances to practice. And then they will feel more confident and they would do a better job in their final exam. And then which you know will give this test more valid. (Instructor 3)

Instructor 5, who used the same syllabus as the researcher, said that she would spend more class time on evaluating sources because that was one aspect of source use that was not taught previously in her sections of English 101C:

> I think if I would want to use this test as a final exam in 101C, I may need to spend more time on like how to use web-based resources properly in writing

> compared to what I had done in my previous 101C I guess…this test is a way to evaluate the skill…So that will also positively influence the things that I teach in my 101C as well….So students will also see the importance to this concept as well because that's I mean one of the skills that's evaluated for their final exam, and everything that's on the exam, they students tend to care more about, right?...If this test would be used in my 101C, then I would have a second thought on this concept, and then um think through how this could be taught better than what we had to be taught previously. And I really didn't think about teaching how to evaluate the credibility of web sources before, but cause the way this test was designed was to let students to freely choose any sources they want to use, right? (Instructor 5)

Instructor 4, who was against adopting the test in his English 101C section, nonetheless thought that test results could provide useful information to the instructor in terms of changes that might need to be made to the syllabus, teaching style, and clarity of the rubric:

> Based on how they [students] performed, I can revise my syllabus or the way I taught, you know how they should cite sources. It will let me know whether they got it or they didn't get it….And also based on what they do, I can inform myself you know I can take a look at the rubrics again to see whether they got it in the sense of whether it was clear to them….it could be for gathering very useful information you know for washback,…looking at my own teaching style and how I communicated to them, and how they also performed, and how you know the student who are coming in the future you know should be handled. (Instructor 4)

Based on the five previous and current English 101C instructors' responses, it seems that the use of the integrated writing test will have positive washback effects on teaching in the English 101C classrooms, particularly with regard to teaching the skills of evaluating and using sources to students.

Controlled Rating Time and Timely Distribution of Score Reports

Research question 7.2 was "How long does it take for raters to rate an essay? How long does it take for raters to rate essays for two course sections?" The answer to this question was necessary to provide backing that can support the assumption that the rating time can be controlled to within a reasonable amount of time so that the score reports can be distributed to

test takers as soon as possible after the test. The answer to research question 7.2 came from the record of time it took second raters to rate each essay.

The average time was 6.94 or around 7 minutes for each essay (range: 3-11 minutes). This means that if an instructor were to teach two sections of English 101C in a semester, he or she would have to rate 40 to 48 final exam essays, meaning that it would take around 280 to 340 minutes to rate the whole batch of essays. This translates to around 5 to 6 hours of rating time. Distribution of score reports would also involve time spent online either uploading individual score report files to Moodle or sending individual emails to students, which would require an additional 2 hours or so. Since instructors must submit course grades by Tuesday afternoon after the week of final exams, instructors usually have at least 4 full days for rating, assuming that the final exam is given on Friday afternoon during finals week. However, exam days and times are assigned by the university every semester and can fall on any weekday during finals week, so it is quite probable that instructors will have more than 4 days to rate essays, send score reports to students, and submit final course grades to the university.

Summary of Results

Table 4.25 summarizes the results according to the research questions and the sources used to answer them.

Table 4.25

Summary of Results According to Research Question

Research Question	Answer (Whether Backing Supports Assumption)	Source

Domain Description	1.1 Domain analysis (skills, knowledge, abilities, and processes) – What are the important skills, knowledge, abilities, and processes needed for source-based academic writing in college courses as identified by experts, syllabi, and textbooks?	Yes	Analysis of English 101C syllabus and textbook and instructor interviews
	1.2 Domain analysis (possible assessment tasks) – What are possible assessment tasks that are representative of the domain of source-based academic writing in college courses as identified by experts, syllabi, and assignment sheets?	Yes	Analysis of English 101C and 150 assignment sheets and instructor interviews
	1.3 Systematic process of task design and modeling – How much did experts think that the web-search-permitted integrated writing test samples important skills and is representative of the domain?	Partially	Instructors thought the test samples important skills taught in English 101C but perhaps not those taught in English 150 or college courses in general.
Evaluation	2.1 Multiple task administration conditions – How did the test takers feel about the test administration conditions (instructions and time limit)?	Yes	Test takers found the instructions clear and the time limit appropriate.
	2.2 Systematic rubric development – What did experts think about the appropriateness of the rating rubric for providing evidence of web-source-based academic writing ability?	Partially	Instructors had many suggestions for improvement of the rubric. The rubric was revised, but not all suggestions were accepted.
	2.3 Rater training and calibration – How much can instructors be trained to avoid bias for or against different groups of students?	Yes	Benchmark essays were provided to raters for self-training. Essays were stripped of identifying information. No significant differences were found in mean essay scores between gender groups and first language groups.

Generalization	3.1 Systematic development of test specification for producing parallel tasks – How much did experts find the test task specification well defined for producing parallel tasks?	Yes	Instructors thought the specification would enable them to produce parallel tasks. More concrete research would strengthen support.
	3.2 Intra-rater reliability – How high is the intra-rater reliability?	Yes	Cronbach's alpha was 0.885, which indicates a good level of reliability.
	3.3 Inter-rater reliability – How high is the inter-rater reliability?	Yes	Cronbach's alpha was 0.770, which indicates an acceptable level of reliability.
Explanation	4.1 Comparative analysis of test task and instructional tasks – How much does the test task reflect instructional tasks in English 101C?	Yes	Instructors thought the test task reflected instructional tasks in English 101C.
	4.2 Comparative analysis of test rubric and rubrics used in courses – How much does the test rubric reflect the rubrics used to evaluate writing in English 101C?	Yes	Instructors thought the test rubric reflected rubrics used in English 101C.
	4.3 Test completion processes and discourse analysis of products – What test-taking processes did test takers follow, and what web-searching behaviors did test takers show? What online language help tools did test takers consult? What relationships are there between test-taking processes and test scores? Do the test scores reflect how well web sources are used in the essays? How do the selection of sources, attribution to sources, and integration of source language relate to scores or differ across score levels?	Partially	Test takers' test-taking processes reflected meaningful construct-relevant activities with regard to use of web sources and online language help options. The material component scores reflected the test takers' skills in using source language in-text. The test scores also reflected test takers' use of references lists. However, the selection of sources was not reflected in the scores.
	4.4 Comparison studies of group differences – How is test performance related to test takers' English learning, writing experience, and internet use?	No	No differences in test scores were found according to test takers' English learning, writing experience, or internet use.

Extrapolation	5.1 Criterion-related evidence – How is test performance related to students' self-assessment of their source-based academic writing ability and students' performance in a post-English 101C course which requires the completion of source-based writing assignments?	Partially	Test performance correlated with test takers' self-assessment of source-based writing ability. However, test performance did not correlate with the final course grades for English 150, although test takers saw correspondence between test content and expectations of future course assignments.
Utilization	6.1 Equal opportunity to learn – How equal did test takers perceive the instruction and preparation they received before the test?	Yes	Test takers thought everyone received equal instruction and preparation before the test.
	6.2 Usefulness, clarity, and interpretability of score descriptors – What did test takers and experts think about the usefulness, clarity, and interpretability of the score descriptors?	Partially	More than half of the test takers and instructors perceived the score descriptors to be useful, clear, and interpretable, but suggestions were made for improvement.
Implication	7.1 Washback studies – What are the washback effects of test use on instruction and learning?	Yes	Many (six out of nine) test takers engaged in meaningful test preparation activities. All instructors predicted positive washback effects on teaching.
	7.2 Controlled rating time and timely distribution of score reports – How long does it take for raters to rate an essay? How long does it take for raters to rate essays for two course sections?	Yes	The average time taken to rate each essay was 6.94 or around 7 minutes (range: 3-11 minutes).

Research question 1.1 asked, "What are the important skills, knowledge, abilities, and processes needed for source-based academic writing in college courses as identified by experts, syllabi, and textbooks?" in order to define and describe the target domain. Based on the analysis of the English 101C syllabus and textbook and instructor interviews, it was possible to extensively outline numerous academic writing skills and processes that are taught in English 101C and receive focus in instruction, including searching for sources, using source language in-

text through summarizing, paraphrasing, and quoting, and adding citations. Therefore, I concluded that yes, the research question can be answered and the assumption can be supported by the data.

Research question 1.2 asked, "What are possible assessment tasks that are representative of the domain of source-based academic writing in college courses as identified by experts and assignment sheets?" in order to define and describe the target domain in terms of task types. Based on the analysis of the English 101C and 150 assignment sheets and instructor interviews, it was possible to come up with representative task types that can be used as the format of the assessment task. Therefore, I concluded that yes, the research question can be answered and the assumption is supported by the data.

Research question 1.3 asked, "How much did experts think that the web-search-permitted integrated writing test samples important skills and is representative of the domain?" to investigate whether an assessment task that requires important skills and is representative of the domain can be simulated. Instructors thought the test samples important skills taught in English 101C but perhaps not those taught in English 150 or college courses in general. Therefore, I concluded that the assumption can be partially supported by the data.

Research question 2.1 asked, "How did the test takers feel about the test administration conditions (instructions and time limit)?" to ensure that task administration conditions are appropriate for providing evidence of targeted language abilities. Multiple task administration conditions were piloted, and based on post-test questionnaires and interviews, test takers found the instructions clear and the time limit appropriate. Therefore, I can conclude that yes, the assumption is supported by the data.

Research question 2.2 asked, “What did experts think about the appropriateness of the rating rubric for providing evidence of web-source-based academic writing ability?” to find out whether the rubric is appropriate for providing evidence of source-based academic writing ability and has been applied as intended. Instructors had many suggestions for improvement of the rubric. With the goal of systematic rubric development, the rubric was revised, but not all suggestions were accepted. Therefore, I concluded that the assumption is partially supported by the data.

Research question 2.3 asked, “How much can instructors be trained to avoid bias for or against different groups of students?” to show that rater training and calibration reduces bias. Benchmark essays were provided to raters, and essays were stripped of identifying information. There were no significant differences in mean essay scores between the two genders and between Chinese and non-Chinese students. Therefore, I concluded that yes, the assumption is supported by the data.

Research question 3.1 asked, “How much did experts find the test task specification well defined for producing parallel tasks?” to evaluate whether task and rating specifications are well defined so that parallel tasks can be created. All five instructors thought that the specification would enable them to produce parallel tasks. I concluded that yes, the assumption is supported by the data, though more concrete research would further strengthen the support.

Research question 3.2 asked, “How high is the intra-rater reliability?” to demonstrate that different ratings by the same instructor are consistent. Cronbach’s alpha was 0.885, which indicates a good level of reliability. Therefore, I concluded that high is yes, the assumption is supported by the data.

Research question 3.3 asked, "How high is the inter-rater reliability?" to show that ratings of different instructors are consistent. Cronbach's alpha was 0.770, which indicates an acceptable level of reliability. Therefore, I concluded that the assumption is supported by the data.

Research question 4.1 asked, "How much does the test task reflect instructional tasks in English 101C?" to ensure that the characteristics of the test correspond closely to those of instructional tasks. Upon comparison of the test task with instructional tasks in English 101C, instructors thought the test task reflected instructional tasks in English 101C. Therefore, I concluded that yes, the assumption is supported by the finding.

Research question 4.2 asked, "How much does the test rubric reflect the rubrics used to evaluate writing in English 101C?" to ensure that the criteria and procedures for evaluating the responses to the test correspond closely to those that instructors have identified as important for assessing performance in other writing tasks in English 101C. Instructors thought, upon comparison, that the test rubric reflected rubrics used in English 101C. Therefore, I concluded that yes, the assumption is supported by the finding.

Research question 4.3 asked, "What test-taking processes did test takers follow, and what web-searching behaviors did test takers show? What online language help tools did test takers consult? What relationships are there between test-taking processes and test scores? Do the test scores reflect how well web sources are used in the essays? How do the selection of sources, attribution to sources, and integration of source language relate to scores or differ across score levels?" to find out whether the linguistic knowledge, processes, and strategies required to successfully complete the test are in keeping with theoretical expectations. Test takers' test-taking processes reflected meaningful and construct-relevant activities with regard to use of web

sources and online language help options. The material component scores reflected the test takers' skills in using source language in-text. The test scores also reflected test takers' use of references lists. However, the selection of sources was not reflected in the test scores. Therefore, I concluded that the assumption is partially supported by the data.

Research question 4.4 asked, "How is test performance related to test takers' English learning, writing experience, and internet use?" to see if test performance varies according to amount and quality of experience in learning English and learning to write from web sources in English. Using post-test questionnaire responses, no differences in test scores were found according to test takers' English learning, writing experience, or internet use. Therefore, I concluded that no, the assumption cannot be supported by the data.

Research question 5.1 asked, "How is test performance related to students' self-assessment of their source-based academic writing ability and students' performance in a post-English 101C course which requires the completion of source-based writing assignments?" to investigate whether performance on the test is related to other criteria of source-based writing ability in the academic context. Test performance correlated with test takers' self-assessment of source-based writing ability. However, test performance did not correlate with future performance in a post-English 101C composition course as measured by the final course grade from English 150, although test takers saw correspondence between test content and expectations of future course assignments. Therefore, I concluded that the assumption is partially supported by the data.

Research question 6.1 asked, "How equal did test takers perceive the instruction and preparation they received before the test?" Test takers who participated in follow-up interviews thought everyone received equal instruction and preparation before the test. Therefore, I

concluded that yes, the assumption that equal instruction and preparation was given to test takers is supported by the data.

Research question 6.2 asked, "What did test takers and experts think about the usefulness, clarity, and interpretability of the score descriptors?" to find out whether the score descriptors are useful, clear, and easy to interpret. According to questionnaires and interviews, more than half of the test takers and instructors perceived the score descriptors to be useful, clear, and interpretable, but suggestions were made for improvement. As a result, the assumption was partially supported by the data.

Research question 7.1 asked, "What are the washback effects of test use on instruction and learning?" to investigate whether the use of the test promotes positive washback on instruction and learning in English 101C. Follow-up questionnaires and interviews revealed that many test takers engaged in meaningful test preparation activities, while all instructors predicted positive washback effects on teaching. Therefore, I concluded that yes, the assumption can be supported by the findings.

Research question 7.2 asked, "How long does it take for raters to rate an essay? How long does it take for raters to rate essays for two course sections?" to ensure that score reports can be distributed to the test takers in a timely manner. The raters spent around 7 minutes on average (range: 3-11 minutes) per essay, which means around 7-8 hours are needed to rate the essays and send the score reports to students in two sections of English 101C. This enables the score reports to be distributed to test takers within at most a week after the administration of the test. Therefore, I concluded that yes, the assumption can be supported by the data.

CHAPTER 5

CONCLUSION

The results and discussion in Chapter 4 become the evidence or backing which supports the interpretive argument presented in Chapter 2. The interpretive argument and evidence are combined and presented together in this chapter as a validity argument for the use of scores from the web-search-permitted and web-source-based integrated writing test. This chapter also discusses implications of the study for validation research in language assessment, recommendations for English 101C instructors, limitations of the study, and suggestions for future research.

A Validity Argument for the Web-Search-Permitted Integrated Writing Test

This validity argument uses supporting evidence collected through this dissertation study to make an argument for the use of scores from a web-search-permitted and web-source-based integrated writing test. Scores from the test are intended to be used as final exam scores in English 101C, an academic writing course for international undergraduate students at Iowa State University. This validity argument is meant to be a stand-alone document, though readers are encouraged to refer to the detailed results presented in Chapter 4 as needed.

The Test

The web-search-permitted and web-source-based integrated writing test is a classroom-based test that can be used as a final exam in English 101C, which is the second of a two-course sequence of ESL academic writing courses for international undergraduate students at Iowa State University. The test prompts test takers to demonstrate skills pertaining to web-source-based

writing, particularly (a) searching for sources online through library databases and search engines, (b) evaluating the credibility of online sources, and (c) using sources in writing by summarizing, paraphrasing, and quoting and by using in-text citations and references lists. The test consists of one essay writing task, which asks test takers to compose an argumentative essay using internet sources that they have searched for and selected during the test. Test takers are also allowed to consult online writing help options, such dictionaries, thesauruses, and grammar checkers. The task tries to draw out test-taking behavior that resembles the processes that would normally be involved in web-source-based writing but within a fairly limited amount of time (two hours). The construct that the test is intended to measure is "web-researching-to-write" or "source-based academic writing ability," with "source" referring specifically to "internet sources."

An Overview of Test Interpretations, Uses, and Consequences

Scores from the integrated writing test have several intended meanings associated with them. The primary semantic meaning that is addressed in this validity argument is the following test score interpretation: The final essay scores are intended to reflect the extent to which students are able to write an argumentative essay while integrating information obtained from internet sources that they have searched for during the test and while making use of online language help options. The interpretations that are going to be made on the basis of the test scores pertain to how well students can demonstrate the source-based writing skills that they have learned and practiced in the English 101C classes.

The policy meanings that are addressed in this validity argument are related to the decisions that will be made based on the test scores and the intended consequences of test use.

The decisions that this validity argument aims to support is the assignment of differential final exam grades as the test is used for the purpose of summative assessment at the end of a semester of ESL academic writing instruction. The final essay score will account for 10% of the final course grade. Therefore, the decisions to be made on the basis of test scores are relatively low-stakes because it is not very likely for a student to fail the class due to bad performance on the test.

There are several intended consequences of the use of the test. Firstly, students will get an idea of how much they have learned and improved after taking the course based on their performance on the test and after seeing the evaluation of results. The instructor, on the other hand, will get an idea of how much students have internalized the skills taught in the course. Furthermore, instructors and students will see the importance of internet literacy and source use in writing in the academic context. Another intended consequence is that test use will promote positive washback on teaching, so that instructors will pay more careful attention to teaching their students how to search for sources, evaluate sources, and incorporate information from sources into their essays.

The Validity Argument

The data that were used to produce backing or evidence that can support the argument for the validity of test interpretation and use include test-taking process data (Camtasia screen capture recordings), test product data (essays), test taker and instructor perception data (post-test student questionnaires and interviews, follow-up student questionnaires and interviews, and instructor interviews), and artifacts (syllabi, textbooks, and assignment sheets). The overall approach to presenting the backing in support of the interpretive argument is to begin with the

first inference (domain description) and go through the entire argument inference by inference. The interpretive argument begins with the domain description and observations of test takers' performance and works its way up the ladder or staircase (Chapelle, 2008, p. 349) to reach the final conclusion and build the whole argument (Figure 8). This is also in accordance with Kane's (2006) analogy of crossing the inference bridges one by one with a warrant or ticket that is supported by backing.

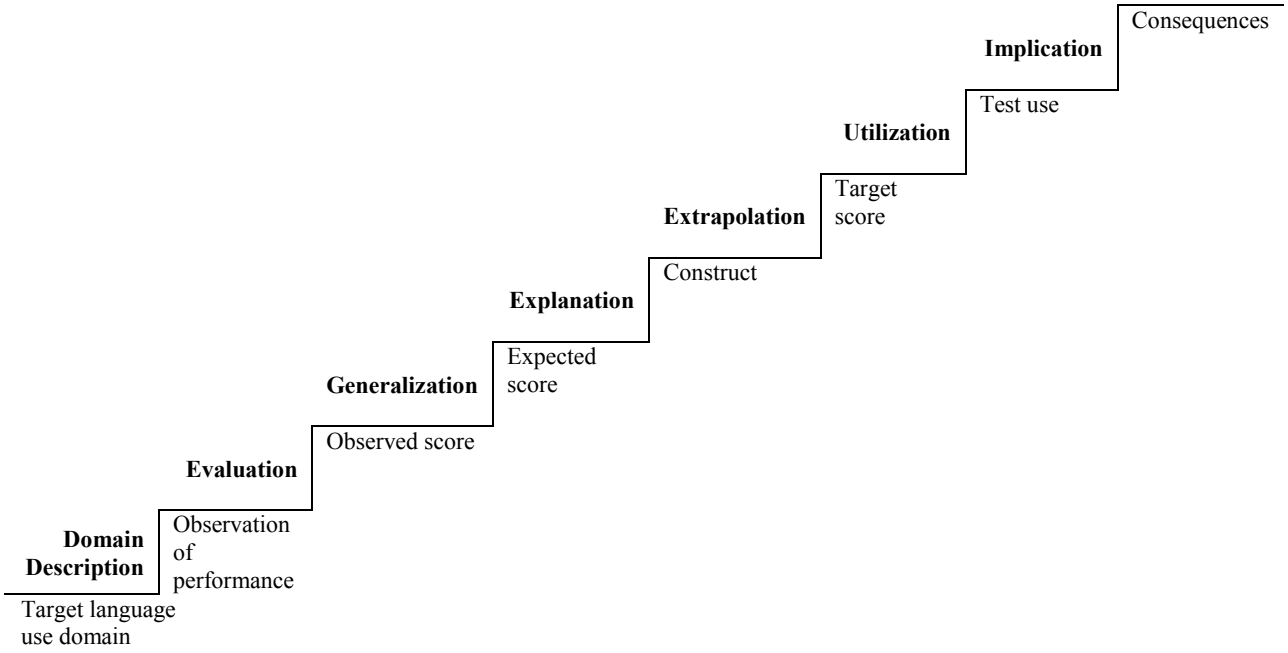

Figure 8. Interpretive argument visualized as a staircase, showing inferences in need of backing (in bold) and grounds/claims (in plain text).

In the validity argument for the use of scores from the web-search-permitted and web-source-based integrated writing test, there are a total of seven inferences and eight grounds/claims. The chain of inferences connects the target language use domain and test scores, which lie at the bottom of the argument, to the intended test use and consequences, which lie at the top of the argument. Each inference connects grounds or data to a claim or an intermediate conclusion. The claim becomes the data for the following inference. It should be noted that the

claims or intermediate conclusions are actually statements, and the expressions in Figure 8 are used as short-hand for the statements.

Domain Description

The first inference in the validity argument is domain description that links the target language use domain to the observation of performance, specifically the essay that the test taker produces. The claim about the observations of performance is that they reflect representative features of the domain of web-source-based academic writing in terms of ability and task. The domain description inference is supported by the warrant that observations of performance on the integrated writing test reveal relevant skills, knowledge, abilities, and processes in situations representative of those in the target domain of web-source-based academic writing in college courses, particularly those knowledge and skills that are outlined in the English 101C syllabus under course goals and taught in the course. As shown in Figure 9, there are three assumptions underlying this warrant: (a) critical English language skills, knowledge, abilities, and processes needed for source-based academic writing in English-medium college classes can be identified; (b) possible assessment tasks that are representative of the domain can be identified; and (c) an assessment task that requires important skills and is representative of the domain can be simulated and systematically developed.

The first two assumptions were supported by backing collected through domain analysis. The methods of expert consensus and analysis of syllabi, textbooks, and assignment sheets helped identify the critical skills, knowledge, abilities, and processes needed in the target domain and possible assessment tasks that can represent the domain. The experts were two previous and current instructors of English 101C. The third assumption was partially supported by backing

related to the systematic process of task design and modeling. The test task was presented to the experts and their comments and feedback on the appropriateness of the task for the purpose of the test were collected. The instructors agreed that the test task represents the domain of web-source-based writing in English 101C.

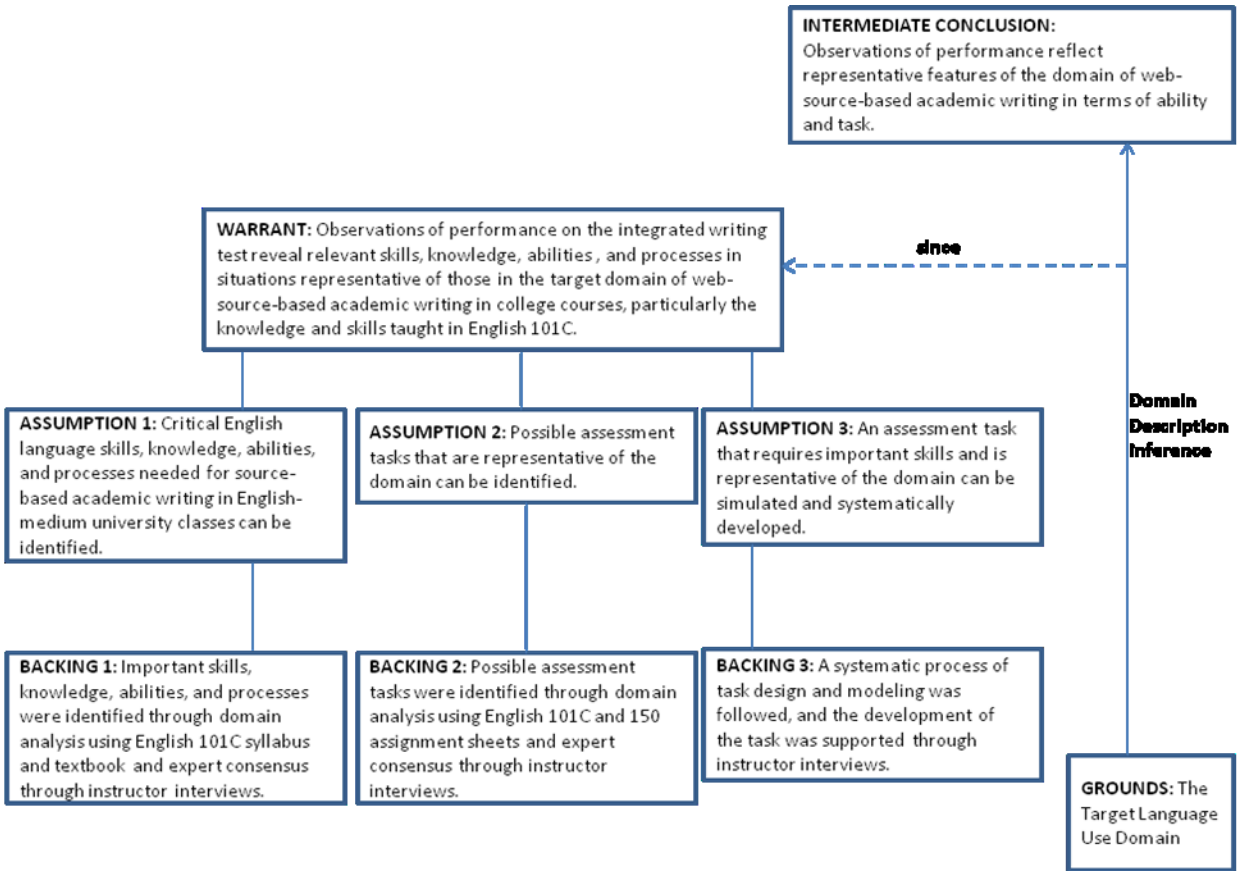

Figure 9. Domain description inference with three assumptions and backing.

Evaluation

The second inference is evaluation, which links the observation of performance to an observed score. The claim about the observed score is that it reflects relevant aspects of the test takers' observed performance. The evaluation inference is supported by the warrant that observations of performance on the integrated writing test are evaluated to provide observed

scores reflective of targeted language abilities. As shown in Figure 10, this warrant is based on three assumptions about task administration conditions and scoring: (a) task administration conditions are appropriate for providing evidence of targeted language abilities; (b) the rubric for scoring essays is appropriate for providing evidence of source-based academic writing ability and has been applied as intended; and (c) instructors can be trained to avoid bias for or against different groups of students.

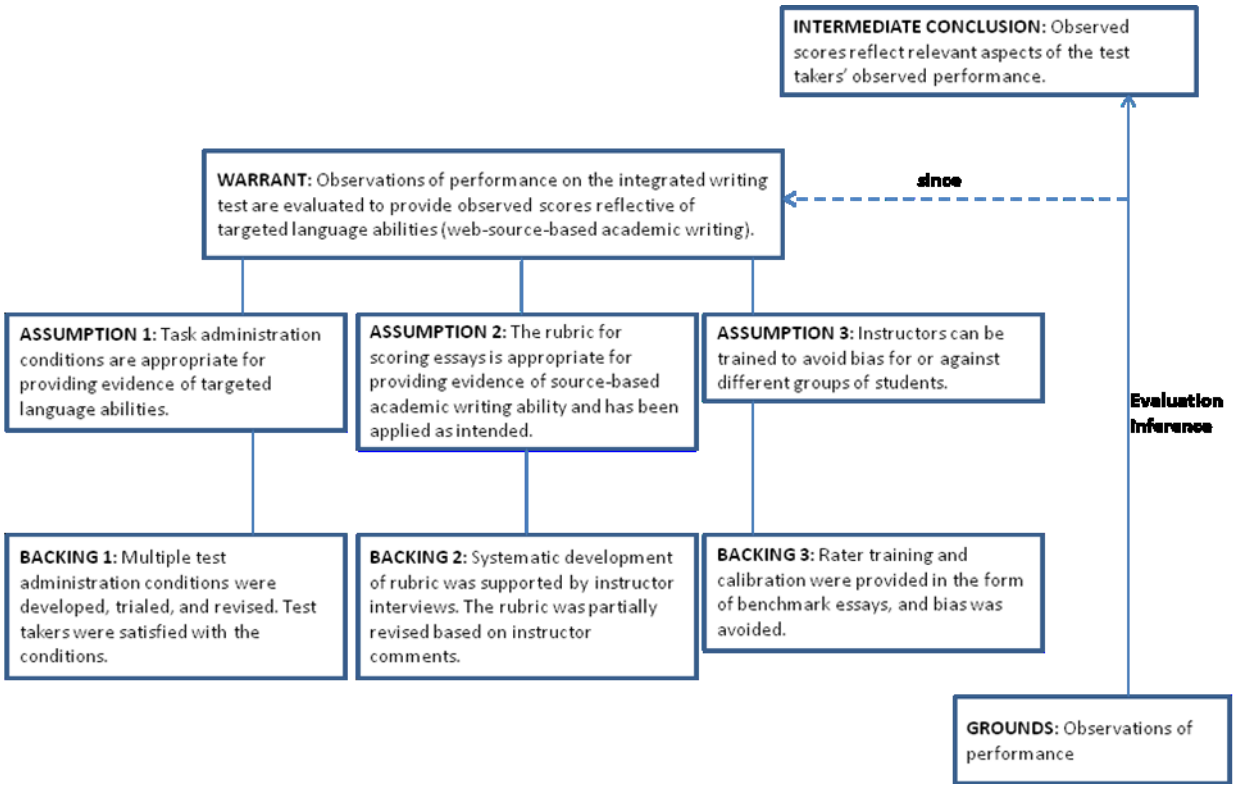

Figure 10. Evaluation inference with three assumptions and backing.

The backing for the first assumption comes from the test development process: multiple task administration conditions were developed, trialed, and revised. Different conditions that were trialed include wording of the prompt, delivery mode of the prompt (Moodle quiz or email), and time limit (90 minutes or 120 minutes). Other backing comes from the test takers' post-test questionnaire responses which confirmed that the test administration conditions were appropriate.

Backing for the second assumption came from systematic rubric development. The rubric was developed and revised based on expert consensus of important, relevant criteria and expert opinion on a draft of the rubric. However, not all comments from the instructors were accepted during the revision process, and therefore, more research is needed on other variations of criteria, score levels, and score descriptors. Lastly, backing for the third assumption came from the provision of rater training and calibration in the form of benchmark essays, blind rating procedure with identifying information removed, and no significant differences in mean essay scores between gender groups and first language groups.

Generalization

The third inference in the validity argument is generalization, which links the observed score to an expected score. The claim about the expected scores is that they reflect what observed scores would be across parallel tasks and within and across raters. The generalization inference is supported by the warrant that observed scores are estimates of expected scores over the relevant parallel versions of tasks and within and across raters. As Figure 11 illustrates, there are three assumptions underlying this warrant: (a) task and rating specifications are well defined so that parallel tasks can be created; (b) different ratings by the same instructor are consistent; and (c) ratings of different instructors are consistent.

The first assumption was supported by the backing that a test specification was systematically developed for the production of parallel tasks. Interviews with five previous and current English 101C instructors showed that all instructors thought they could produce parallel tasks from the specification, but more concrete research is needed. The second and third assumptions were supported by examination of intra-rater reliability and inter-rater reliability,

respectively. The intra-rater reliability estimate was good, while the inter-rater reliability estimate was acceptable.

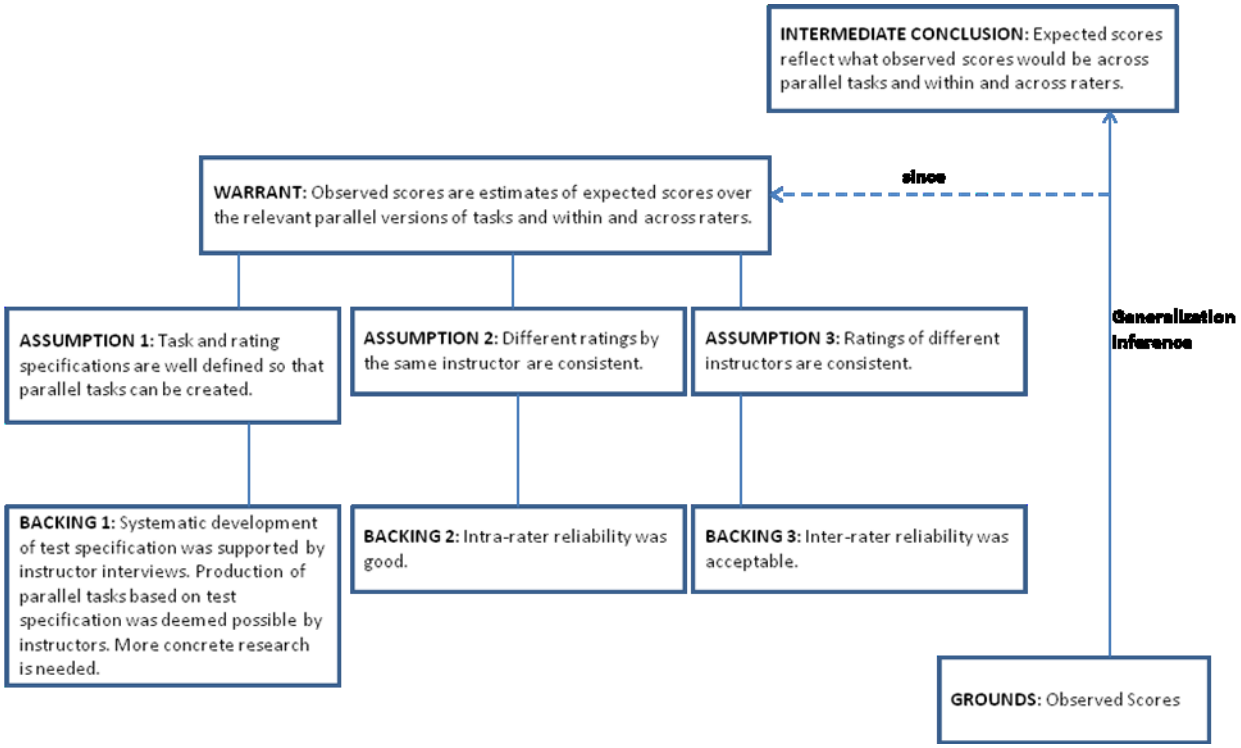

Figure 11. Generalization inference with three assumptions and backing.

Explanation

The fourth inference is explanation, which links the expected score to the construct. The claim about the construct is that it is delimited based on the English 101C syllabus and the teaching/learning activities that occurred in the course. The explanation inference is supported by the warrant that expected scores are attributed to a construct of web-source-based academic writing ability, which is defined by the English 101C syllabus and the teaching/learning activities in the class. In other words, the interpretations about the students' ability to search for and select internet sources and incorporate information from the sources in an argumentative essay are meaningful with respect to the English 101C syllabus and the teaching/learning activities in the

class. As displayed in Figure 12, there are three assumptions underlying this warrant: (a) the characteristics of the integrated writing test correspond closely to those of instructional tasks; (b) the criteria and procedures for evaluating the responses to the integrated writing test correspond closely to those that instructors have identified as important for assessing performance in other writing tasks in the instructional setting; and (c) the linguistic knowledge, processes, and strategies required to successfully complete tasks vary in keeping with theoretical expectations.

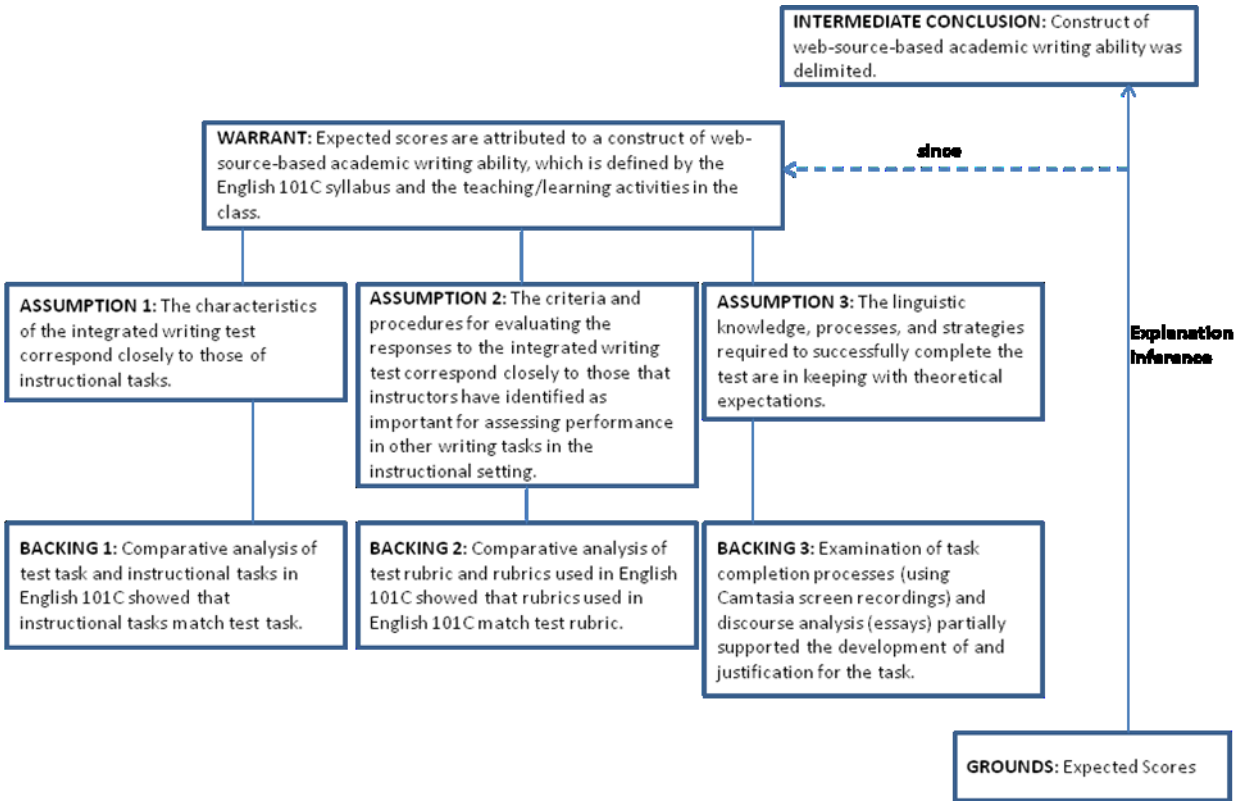

Figure 12. Explanation inference with three assumptions and backing.

Backing for the first assumption came from comparative analysis of the test task and instructional tasks, while for the second assumption, backing came from comparative analysis of the test rubric and rubrics used for essay assignments in English 101C. Instructors thought that the test task matches the instructional tasks and that the test rubric matches the rubrics used in English 101C. The third assumption was partially supported by the backing that examination of

task completion processes using Camtasia screen recordings and discourse analysis of the essays partially supported the development of and justification for the test task. The test takers showed test-taking behaviors and strategies that reflect the construct, while the essays displayed web-source-based academic writing skills. The test scores reflected the varying skills of the test takers in integrating source language and attributing to sources in-text. However, selection of credible sources was not captured by the test scores, which points to a need to further revise the rating rubric.

Extrapolation

Extrapolation, which is the fifth inference in the validity argument, links the construct to the target score. The claim about the target scores is that they represent performance in an academic writing context. The warrant that supports the extrapolation inference is that the construct of source-based academic writing as assessed by the integrated writing test accounts for the quality of source-based academic writing performance in college courses, particularly with regard to those skills that are identified in the English 101C syllabus under course goals and taught in the class. As illustrated in Figure 13, the assumption underlying this warrant is that performance on the test is related to other criteria of source-based writing ability in the academic context.

Backing was sought from criterion-related evidence which examine the relationships between test performance and (a) students' self-assessment of their source-based academic writing ability and (b) students' performance in a post-English 101C course which requires the completion of source-based writing assignments. Results indicated a significant positive relationship between test scores and students' self-assessment of their source-based academic

writing ability. A relationship between test scores and students' final course grades from a post-English 101C course (English 150) was not found, but according to follow-up interviews with nine test takers, test content and requirements extrapolated to future course content and requirements.

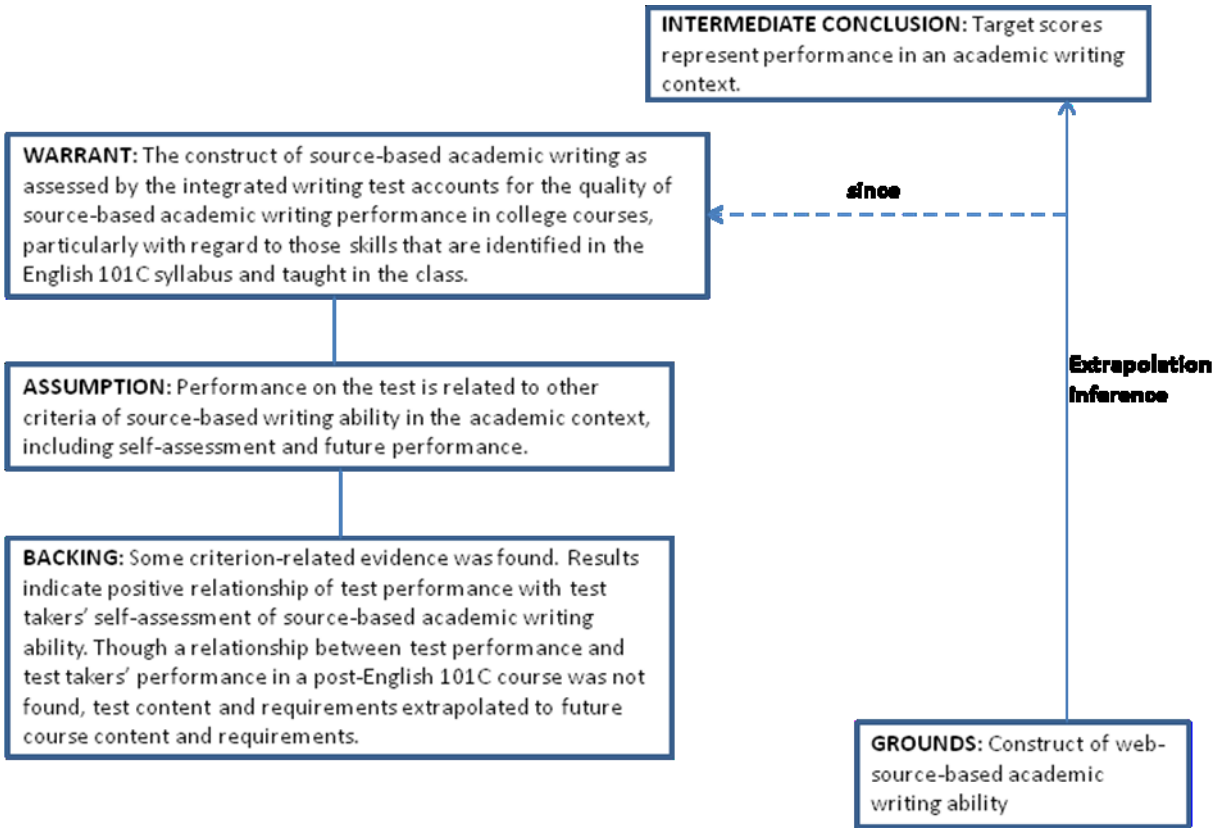

Figure 13. Extrapolation inference with one assumption and backing.

Utilization

The sixth inference in this validity argument is utilization, which links the target score to test use. The claim about test use is that meaningful and equitable decisions are made based on target scores to assign final exam grades and to guide English 101C instruction. The utilization inference is supported by the warrant that estimates of the ability to search for and select internet sources and incorporate information from the sources in an argumentative essay, which are

obtained from the integrated writing test, are useful for making decisions about final exam grades and appropriate curricula for students in English 101C. Furthermore, the summative decisions that are made about students' progress reflect relevant existing educational and societal values and relevant university regulations and are equitable for the 101C students. These decisions are made by the 101C instructors. As Figure 14 illustrates, two assumptions underlie the warrant: (a) students have equal opportunities to learn or acquire the ability to write from internet sources in English 101C; and (b) the test scores provide useful and meaningful information to the students and instructors regarding students' source-based writing abilities, and the meaning of test scores is clearly interpretable by test takers and instructors.

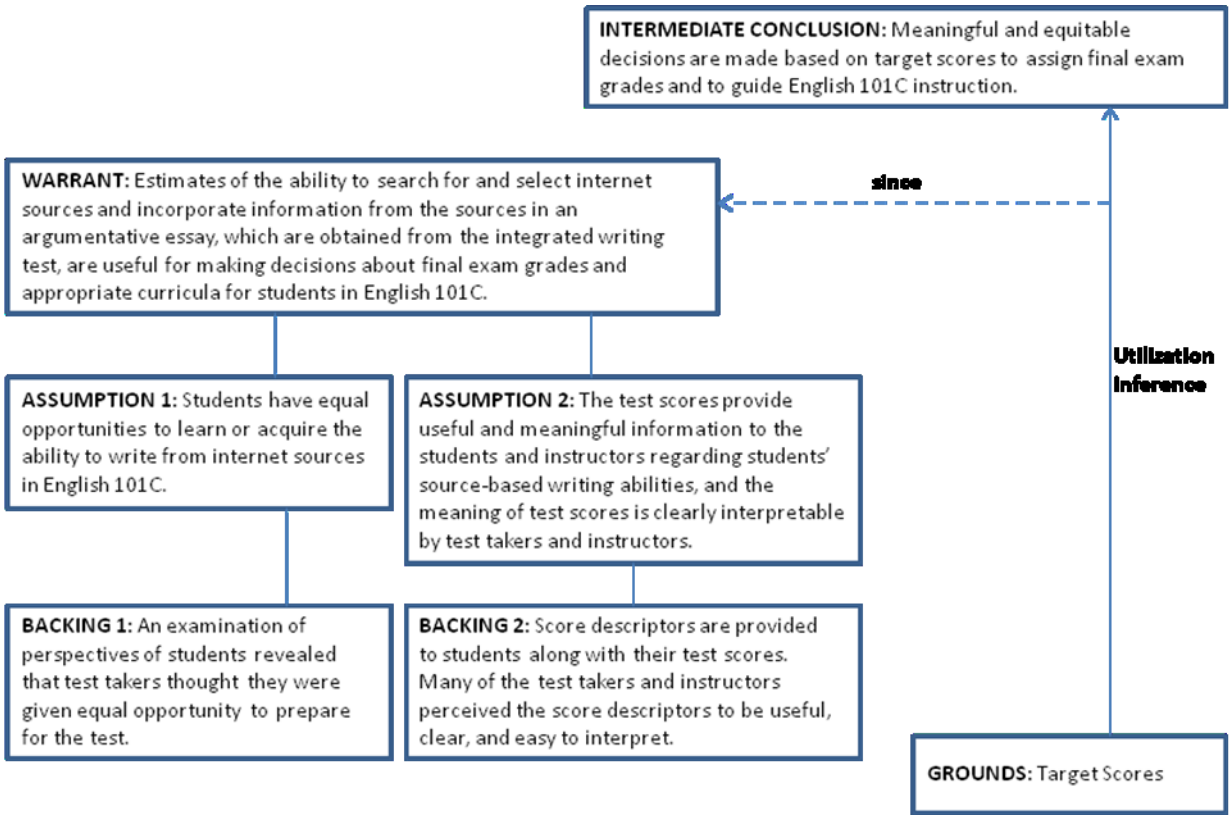

Figure 14. Utilization inference with two assumptions and backing.

Backing for the first assumption came from perspectives of students. Follow-up interviews revealed that all test takers thought they were given equal opportunity to prepare for

the test by receiving instruction in English 101C on how to write an essay and how to search for and use web sources. The second assumption was supported by the backing that score descriptors are provided to students along with their test score. Further backing for the second assumption came from students' and instructors' perspectives on the usefulness, clarity, and interpretability of the descriptors. Many test takers and instructors perceived the score descriptors to be useful, clear, and easy to interpret. However, some test takers and instructors made suggestions for the descriptors which prompt further revision.

Implication

The seventh and final inference is implication, which links the test use to the intended consequences. The claim about the consequences is that they are brought about by the use of scores from the test as intended, including positive washback effects on English 101C instruction and learning. The warrant that supports the implication inference is that the consequences of using the integrated writing test and the decisions that are made are beneficial to the students (test takers), the English 101C teachers, English 150 teachers who will teach these students in the following semester, and instructors of academic courses at ISU who will encounter these students in their classes. The stakeholders, therefore, are (a) students in the English 101C classes, (b) the English 101C instructors, and (c) other instructors who will teach the students in future semesters. As shown in Figure 15, the two assumptions that underlie the warrant are that (a) the test has a positive influence on how academic writing is learned and taught, that is, test use promotes positive washback effects on English 101C; and (b) score reports are distributed to students in a timely manner.

The backing for the first assumption was gathered from a washback study using instructor interviews, follow-up student questionnaires, and follow-up student interviews, while the backing for the second assumption came from the rating sessions by measuring and recording the time it took to rate each essay. The washback study showed that all instructors predicted positive washback effects of test use on teaching, while many test takers engaged in meaningful test preparation activities. As for the second backing, rating time was manageable and allowed for timely distribution of score reports to test takers.

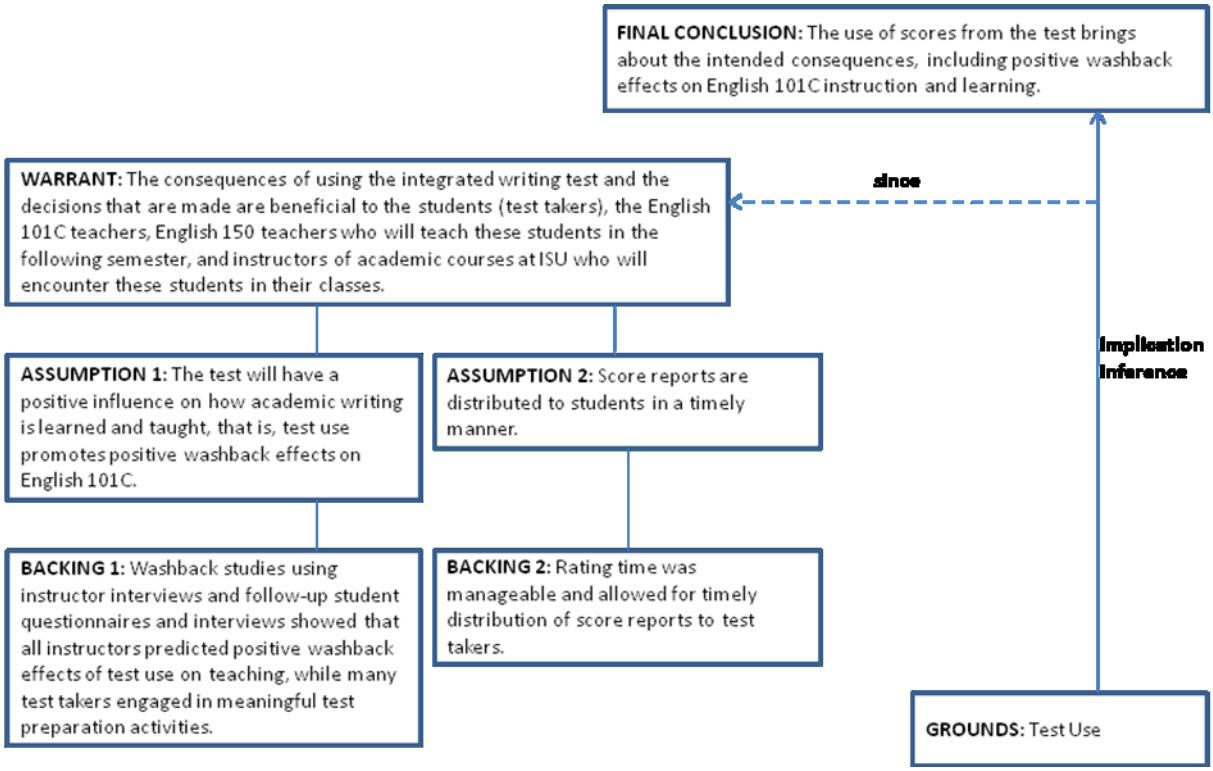

Figure 15. Implication inference with two assumptions and backing.

Table 5.1 summarizes the inferences, warrants, assumptions, and backing in the validity argument for the use of scores from the web-search-permitted and web-source-based integrated writing test.

Table 5.1

Validity Argument for the Web-Search-Permitted and Web-Source-Based Integrated Writing Test

Inference in the Interpretive Argument	Warrant Licensing the Inference	Assumptions Underlying Warrant	Backing to Support Assumption
Domain description (Target language use domain → observation of performance (essay))	Observations of performance on the integrated writing test reveal relevant knowledge, skills, abilities, and processes in situations representative of those in the target domain of web-source-based academic writing in college courses, particularly the knowledge and skills taught in English 101C.	1. Critical English language skills, knowledge, abilities, and processes needed for source-based academic writing in English-medium university classes can be identified. 2. Possible assessment tasks that are representative of the domain can be identified. 3. An assessment task that requires important skills and is representative of the domain can be simulated and systematically developed.	1. Important skills, knowledge, abilities, and processes were identified through domain analysis using English 101C syllabus and textbook and expert consensus through instructor interviews. 2. Possible assessment tasks were identified through domain analysis using English 101C and 150 assignment sheets and expert consensus through instructor interviews. 3. A systematic process of task design and modeling was followed, and the development of the task was supported through instructor interviews.
Evaluation (Observation of performance → observed score)	Observations of performance on the integrated writing test are evaluated to provide observed scores reflective of targeted language abilities (web-source-based academic writing).	1. Task administration conditions are appropriate for providing evidence of targeted language abilities. 2. The rubric for scoring essays is appropriate for providing evidence of source-based academic writing ability and has been applied as intended. 3. Instructors can be trained to avoid bias for or against different groups of students.	1. Multiple test administration conditions were developed, trialed, and revised. Test takers were satisfied with the conditions. 2. Systematic development of rubric was supported by instructor interviews. The rubric was revised based on instructor comments, although not all comments were accepted. 3. Rater training and calibration were provided, and bias was avoided.
Generalization (Observed score → expected/universe score)	Observed scores are estimates of expected scores over the relevant parallel versions of tasks and within and across raters.	1. Task and rating specifications are well defined so that parallel tasks can be created. 2. Different ratings by the same instructor are consistent. 3. Ratings of different instructors are consistent.	1. Systematic development of test specification was supported by instructor interviews. Production of parallel tasks based on test specification was deemed possible by instructors. More concrete research is needed. 2. Intra-rater reliability was good. 3. Inter-rater reliability was acceptable.
Explanation (Expected/universe score → construct)	Expected scores are attributed to a construct of web-source-based academic writing ability, which is defined by the English 101C syllabus and the teaching/learning activities in the class.	1. The characteristics of the integrated writing test correspond closely to those of instructional tasks. 2. The criteria and procedures for evaluating the responses to the integrated writing test correspond closely to those that instructors have identified	1. Comparative analysis of test task and instructional tasks in English 101C showed that instructional tasks match test task. 2. Comparative analysis of test rubric and rubrics used in English 101C showed that rubrics used in English 101C

	The interpretations about the students' ability to search for and select internet sources and incorporate information from the sources in an argumentative essay are meaningful with respect to the English 101C teaching syllabus and the teaching/learning activities in the class.	as important for assessing performance in other writing tasks in the instructional setting. 3. The linguistic knowledge, processes, and strategies required to successfully complete the test are in keeping with theoretical expectations.	match test rubric. 3. Examination of task completion processes (using screen recordings and post-test interviews) and discourse analysis (essays) supported the development of and justification for the task. Processes and products reflected the construct of web-researching-to-write. Scores mostly reflected the construct, but selection of credible sources was not captured by the test scores, which points to a need to further revise the rating rubric.
Extrapolation (Construct → target score)	The construct of web-source-based academic writing as assessed by the integrated writing test accounts for the quality of web-source-based academic writing performance in college courses, particularly with regard to those skills taught in English 101C.	1. Performance on the test is related to other criteria of source-based writing ability in the academic context including self-assessment and future performance.	1. Some criterion-related evidence was found. An examination of the relationship between test performance and students' self-assessment of their own source-based academic writing ability showed that self-assessment was related to test scores. An examination of the relationship between test performance and students' performance in a post-English 101C course revealed that future performance is not related to test scores, although test content and requirements extrapolated to future course content and requirements.
Utilization (Target score → test use/decision)	Estimates of the ability to search for and select internet sources and incorporate information from the sources in an argumentative essay, which are obtained from the integrated writing test, are useful for making decisions about final exam grades and appropriate curricula for students in English 101C. The summative decisions that are made about students' progress reflect relevant existing educational and societal values and relevant	1. Students have equal opportunities to learn or acquire the ability to write from internet sources in English 101C. 2. The test scores provide useful and meaningful information to the students and instructors regarding students' source-based writing abilities, and the meaning of test scores is clearly interpretable by test takers and instructors.	1. An examination of perspectives of students revealed that test takers thought they were given equal opportunity to prepare for the test. 2. Score descriptors are provided to students along with their test scores. Many test takers and instructors perceived the score descriptors to be useful, clear, and easy to interpret, but some made suggestions which prompt further revision.

	university regulations and are equitable for the 101C students. These decisions will be made by the 101C instructors.		
Implication (Test use → consequences)	The consequences of using the integrated writing test and the decisions that are made are beneficial to the students (test takers), the English 101C teachers, English 150 teachers who will teach these students in the following semester, and instructors of academic courses at ISU who will encounter these students in their classes. Stakeholders: 1) Students in the English 101C classes 2) The English 101C instructors 3) Other instructors who will teach the students in future semesters	1. The test will have a positive influence on how academic writing is learned and taught, that is, test use promotes positive washback effects on English 101C. 2. Score reports are distributed to students in a timely manner.	1. Washback studies using instructor interviews and follow-up student questionnaires and interviews showed that all instructors predicted positive washback effects of test use on teaching, while many test takers engaged in meaningful test preparation activities. 2. Rating time was manageable and allowed for timely distribution of score reports to test takers.

Implications, Recommendations, Limitations, and Suggestions

The implications of the study for the field of language assessment, particularly validation research, are as follows. First, this study showed that a validity argument can be constructed for small-scale low-stakes language tests. Possibly because it has not been very long since Kane's argument-based approach to validation was applied to language assessment, some people tend to think that validity arguments are built exclusively for large-scale testing and that the approach is irrelevant to small-scale classroom-based assessment. Therefore, the validity argument presented in this study is meaningful as being one of the first to be constructed for a low-stakes classroom-based test. It has illustrated the applicability of the argument-based approach to classroom testing

and has shown that the core elements of the approach can be shared across all types of assessment. It is hoped that my validity argument will become a useful example for others in the field who wish to construct validity arguments for their own language testing contexts.

Second, the test used in this study is one of the first of its kind to be added to the literature in language assessment. Many recently published studies have dealt with integrated writing tests that provide test takers with one or more pre-selected listening and/or reading text(s), but the web-search-permitted and web-source-based test used in this study opens up the possibility of allowing test takers the freedom (and perhaps the burden) of searching for and selecting sources online by themselves. Also, test takers were given the freedom to utilize online language help options during the test. This study has shown the feasibility of these arrangements in a test situation, as most test takers could successfully handle the added challenge and all test takers made effective use of help options to improve their writing. These test processes can be considered more authentic in reflecting what writers actually do when writing in non-test situations. It is hoped that the new integrated assessment task introduced in this study can be used in the future for both teaching and research purposes.

There are also recommendations for future English 101C instructors based on the findings of this study. First of all, if the test is going to be used operationally across multiple English 101C sections, it would be important to create multiple parallel prompts every semester, so that each section of English 101C is given a different prompt and there are no test security issues with students exchanging information about topics across sections. This point pertains to the administration conditions of the test, i.e., that the test elicits the intended performance rather than pre-prepared material that students might bring to the test session.

Secondly, as suggested by several test takers and instructors interviewed in the study, it is important to present the rating rubric to test takers before the test and clear any confusion over the meanings of terms in the score descriptors. This will help test takers become more prepared and aware of the expectations and requirements of the test. Another procedure in the rating process that can make the score report more useful for the test takers is to highlight specific bulleted points within the boxes in the rubric. This could clarify to the test takers the reasons for their essays receiving a certain score level for each criterion and what specific aspects need improvement, thereby personalizing the feedback provided by the marked rating rubric.

Thirdly, since the analysis of the test essays showed that not all test takers who copied from sources were penalized, instructors are recommended to run students' test essays through plagiarism detection websites such as Turnitin.com or iThenticate, provided that subscriptions to such websites are available at their institutions. Another effective method is to copy and paste suspect sentences into a search engine such as Google to see if there are exact matches. By taking these measures while rating the essays, instructors can ensure that test scores more accurately reflect test takers' abilities.

Certainly there are limitations to the study. The first limitation is that the test was used in three sections of English 101C that were taught by the researcher. Although this resulted in homogeny of instructional materials and in-class activities, it may be difficult to generalize the findings of this study to other sections of English 101C or to other academic writing courses for international undergraduate students in the US. The test could have been used in multiple sections of English 101C taught by numerous instructors to increase the generalizability of the findings.

The second limitation lies in the rating rubric used in the study. The rating rubric, although fairly clear for the instructors and raters, was perhaps too linguistically complex for the students; hence a proposal was made to develop a student version of the rating rubric/score report with simplified score descriptors. The rating rubrics could have been further improved by taking the advice of Galaczi, Lim, and Khabbazbashi (2013) who suggest creating a glossary of terminology for rubrics and clarifying relative words in the rubric such as "comparable" and "slightly." This would have reduced any confusion for the instructors who were rating the essays and would have also increased the reliability of rating.

In light of the limitations of the current study and also the backing that is additionally needed to support the assumptions in the validity argument, there are several suggestions for future research. The most pressing, according to the validity argument, is to create one or more additional rating rubrics and compare them to the current version. The new version would have a separate criterion or criteria for web source use and citation as was suggested by a few instructors in the current study. The findings from this research would add support to the evaluation, explanation, and utilization inferences in the validity argument. This future research would also have much value for investigating rater behavior and performance to find out whether raters do or can separate source use from general development of content.

Secondly, as suggested in discussion of the generalization inference, parallel versions of the test task could be developed and used by multiple instructors. In fact, a generalizability study could look at the prompt effect and rater effect at the same time and add support to the generalization inference.

Thirdly, more objective data should be collected to strengthen the extrapolation inference. Perhaps the source-based writing assignments that students completed in courses post-English

101C can be collected and be graded again by trained raters so that the ratings are comparable to the test scores. Instructors' judgments about students' performance on source-based writing assignments can also be collected as another measure to compare with students' performance on the test.

Fourthly, to further strengthen the implication inference, additional methods can be used for the washback study, including observations and reflective teaching journals. Observations of students' test preparation activities in-class and self-reported accounts of whether and how instructors tried to prepare students for the final exam would provide a clearer picture of the washback effects of test use.

Finally, Plakans (n.d.) points out many unresolved issues with integrated assessment tasks, including topic effect, task effect, rater reliability, construct definition (as integrated tests muddy what you are trying to measure, there is a need to find out how the skills overlap and support each other), challenge for test takers to know what to do even with directions, borrowing chunks of language from sources that can be a problem for raters, and interpreting the scores. All of these are possible avenues of further research that are worth pursuing, the results of which would be valuable in adding to the combined body of knowledge on the validity of using and interpreting scores from integrated language assessment tasks.

REFERENCES

Abasi, A. R., & Graves, B. (2008). Academic literacy and plagiarism: Conversations with international graduate students and disciplinary professors. *Journal of English for Academic Purposes*, *7*, 221-233.

AERA, APA, & NCME. (1999). Chapter 1. Validity. *Standards for educational and psychological testing* (pp. 9-24). Washington, DC: AERA.

Ali, R., & Katz, I. R. (2010). *Information and communication technology literacy: What do businesses expect and what do business schools teach?* (ETS RR-10-17). Princeton, NJ: Educational Testing Service.

Asención, Y. (2004). Validation of reading-to-write assessment tasks performed by second language learners. (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, AZ.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Barks, D., & Watts, P. (2001). Textual borrowing strategies for graduate-level ESL writers. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 246-267). Ann Arbor, MI: The University of Michigan Press.

Bensoussan, M., Sim, D., & Weiss, R. (1981). The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. *Biannual Conference of the International Association of Applied Linguists* (ERIC Document No. ED 232 436).

Biddix, J. P., Chung, C. J., & Park, H. W. (2011). Convenience or credibility? A study of college student online research behaviors. *Internet and Higher Education*, *14*, 175-182.

Blattner, N. H., & Frazier, C. L. (2002). Developing a performance-based assessment of students' critical thinking skills. *Assessing Writing*, *8*, 47-64.

Bloch, J. (2001). Plagiarism and the ESL student: From printed to electronic texts. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 209-228). Ann Arbor, MI: The University of Michigan Press.

Bloch, J. (2009). The design of an online concordancing program for teaching about reporting verbs. *Language Learning & Technology*, *13*(1), 59-78.

Briggs, D. C. (2004). Comment: Making an argument for design validity before interpretive validity. *Measurement: Interdisciplinary Research and Perspectives*, *2*(3), 171-174.

Brown, C. A., Dickson, R., Humphreys, A., McQuillan, V., & Smears, E. (2008). Promoting academic writing/referencing skills: Outcome of an undergraduate e-learning pilot project. *British Journal of Educational Technology*, *39*(1), 140-156.

Burton, V. T., & Chadwick, S. A. (2000). Investigating the practices of student researchers: Patterns of use and criteria for use of internet and library sources. *Computers and Composition*, *17*, 309-328.

Carson, J. G. (2001). A task analysis of reading and writing in academic contexts. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 48-83). Ann Arbor, MI: The University of Michigan Press.

Chan, S. H. C. (2011). *Demonstrating cognitive validity and face validity of PTE Academic writing items* Summarize Written Text *and* Write Essay. Research Note. Pearson Education Ltd. Retrieved from http://pearsonpte.com/research/Documents/RN_DemonstratingCognitiveAndFaceValidityOfPTEAcademicWritingItems_2011.pdf

Chapelle, C. A. (2008). Chapter 9. The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language*^{TM} (pp. 319-352). New York: Routledge.

Chapelle, C. A. (2011). Chapter 43. Validation in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning: Volume II* (pp. 717-730). New York: Routledge.

Chapelle, C. A. (2012a). Validity argument for language assessment: The framework is simple… *Language Testing*, *29*(1), 19-27.

Chapelle, C. A. (2012b, April). Values, computer technology and validation in language assessment. Samuel J. Messick Memorial Lecture. LTRC 2012, Princeton, NJ.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*^{TM}. New York: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3–13.

Christianson, K. (1997). Dictionary use by EFL writers: What really happens? *Journal of Second Language Writing*, *6*(1), 23-43.

Corbett, P. (2010). What about the "Google effect"? Improving the library research habits of first-year composition students. *Teaching English in the Two-Year College*, *38*(3), 265-277.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks: Sage Publications.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5-43.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven: Yale University Press.

Davis, M. (2013). The development of source use by international postgraduate students. *Journal of English for Academic Purposes*, *12*, 125-135.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*, 140-150.

Department of English, Iowa State University. (2012). *ISUComm Foundation Courses Student Guide for English 150 and 250, 2012 – 2013*. Ames: Department of English, Iowa State University.

Dovey, T. (2010). Facilitating writing from sources: A focus on both process and product. *Journal of English for Academic Purposes*, *9*, 45-60.

Dudley-Evans, T. (2002). The teaching of the academic essay: Is a genre approach possible? In A. M. Johns (Ed.), *Genre in the classroom: Multiple perspectives* (pp. 225-235). Mahwah, NJ: Lawrence Erlbaum Associates.

East, M. (2006). The impact of bilingual dictionaries on lexical sophistication and lexical accuracy in tests of L2 writing proficiency: A quantitative analysis. *Assessing Writing*, *11*, 179-197.

Eckel, E. J. (2011). Textual appropriation in engineering master's theses: A preliminary study. *Sci Eng Ethics*, *17*, 469-483.

Erling, E. J., & Richardson, J. T. E. (2010). Measuring the Academic Skills of University Students: Evaluation of a diagnostic procedure. *Assessing Writing*, *15*, 177-195.

Esmaeili, H. (2002). Reading-to-write reading and writing tasks and ESL students' reading and writing performance in an English language test. *Canadian Modern Language Review*, *58*, 599-622.

Faigley, L. (2006). *The brief penguin handbook*. Upper Saddle River, NJ: Pearson Education, Inc.

Flowerdew, J., & Li, Y. (2007). Plagiarism and second language writing in an electronic age. *Annual Review of Applied Linguistics*, *27*, 161-183.

Galaczi, E., Lim, G., & Khabbazbashi, N. (2013, July). Rating scale development and use: The rater perspective. Paper presented at Language Testing Research Colloquium 2013, Seoul, Korea.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*(4), 507-531.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, *15*, 100-117.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spann Fellow Working Papers in Second or Foreign Language Assessment*, *7*, 47-84.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly*, *10*, 9-27.

Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal*, *63*(4), 363-372.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218-238.

Harvey, K., & Yuill, D. (2007). A study of the use of a monolingual pedagogical dictionary by learners of English engaged in writing. *Applied Linguistics*, *18*(3), 253-278.

Harris, R. A. (2011). *Using sources effectively: Strengthening your writing and avoiding plagiarism* (3rd ed.). Glendale, CA: Pyrczak Publishing

Helms-Park, R., & Stapleton, P. (2006). How the views of faculty can inform undergraduate Web-based research: Implications for academic writing. *Computers and Composition*, *23*, 444-461.

Helms-Park, R., Radia, P., & Stapleton, P. (2007). A preliminary assessment of Google Scholar as a source of EAP students' research materials. *Internet and Higher Education*, *10*, 65-76.

Hirvela, A. (2004). *Connecting reading and writing in second language writing instruction*. Ann Arbor, MI: The University of Michigan Press.

Hirvela, A., & Du, Q. (2013). "Why am I paraphrasing?": Undergraduate ESL writers' engagement with source-based academic writing and reading. *Journal of English for Academic Purposes*, *12*(2), 87-98.

Hohlfeld, T. M., Ritzhaupt, A. D., & Barron, A. E. (2010). Development and validation of the Student Tool for Technology Literacy (ST^2L). *Journal of Research on Technology in Education*, *42*(4), 361-389.

International ICT Literacy Panel. (2002). *Digital transformation: A framework for ICT literacy*. Princeton, NJ: Educational Testing Service.

Jones, S. (2002). The Internet goes to college: How students are living in the future with today's technology. Retrieved from the Pew Internet and American Life Project Web site: http://www.pewinternet.org/Reports/2002/The-Internet-Goes-to-College.aspx

Jones, S., Johnson-Yale, C., Millermaier, S., & Perez, F. S. (2008). Academic work, the Internet and U.S. college students. *The Internet and Higher Education*, *11*(3–4), 165−177.

Jordan, R. R. (1997). *English for academic purposes: A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527-535.

Kane, M. (2006). Validation. In R. Brennen. (Ed.), *Educational measurement* (4th ed.) (pp 17-64). Westport, CT: Greenwood Publishing.

Kane, M. (2012). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3-17.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Katz, I. R. (2007). Testing information literacy in digital environments: ETS's iSkills Assessment. *Information Technology and Libraries*, *26*(3), 3–12.

Katz, I. R., Elliot, N., Attali, Y., Scharf, D., Powers, D., Huey, H., Joshi, K., & Briller, V. (2008). *The assessment of information literacy: A case study* (ETS RR-08-33). Princeton, NJ: Educational Testing Service.

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, *14*(1), 28-44.

Kim, J. Y. (2008). Development and validation of an ESL diagnostic reading-to-write test: An effect-driven approach. (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Kim, S. H. (2009). For students, by students: Creating wikis to promote integrity in academic writing and citation. *SLW and CALL Perspectives*, *1*(1). Retrieved from http://www.tesol.org/NewsletterSite/view.asp?nid=3124

Lee, H.-K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Teaching*, *24*(3), 307-330.

Lee, Y.-W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing*, *7*(4), 353-385.

Li, Y. (2013). Three ESL students writing a policy paper assignment: An activity-analytic perspective. *Journal of English for Academic Purposes*, *12*, 73-86.

Lowe, G. S., & McAuley, J. (2000). *Information and communication technology literacy assessment framework*. Unpublished report. Adult Literacy and Lifeskills Survey.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mansourizadeh, K., Ahmad, U. K. (2011). Citation practices among non-native expert and novice scientific writers. *Journal of English for Academic Purposes*, *10*, 152-161.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: Macmillan Publishing Co.

Nesi, H., & Meara, P. (1991). How using dictionaries affects performance in multiple-choice EFL tests. *Reading in a Foreign Language*, *8*(1), 631-643.

Niedbala, M. A., & Fogleman, J. (2010). Taking library 2.0 to the next level: Using a course wiki for teaching information literacy to honors students. *Journal of Library Administration*, *50*, 867-882.

Ohkubo, N. (2009). Validating the integrated writing task of the TOEFL internet-based test (iBT): Linguistic analysis of test takers' use of input material. *Melbourne Papers in Language Testing*, *14*(1), 1-31.

Pecorari, D. (2001). Plagiarism and international students: How the English-speaking university responds. In D. Belcher & A. Hirvela (Eds.), *Linking literacies: Perspectives on L2 reading-writing connections* (pp. 229-245). Ann Arbor, MI: The University of Michigan Press.

Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing*, *12*, 317-345.

Petrić, B., & Harwood, N. (2013). Task requirements, task representation, and self-reported citation functions: An exploratory study of a successful L2 student's writing. *Journal of English for Academic Purposes*, *12*, 110-124.

Plakans, L. (n.d.). 11. Integrated assessment. In G. Fulcher & R. Thrasher. (Eds.), *Language testing videos*. In association with ILTA. Retrieved from http://languagetesting.info/video/main.html

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*, 111-129.

Plakans, L. (2009a). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561-587.

Plakans, L. (2009b). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes, 8*, 252-266.

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly*, *44*(1), 185-194.

Priemer, B., & Ploog, M. (2007). The influence of text production on learning with the Internet. *British Journal of Educational Technology*, *38*(4), 613-622.

Ruiz-Funes, M. (1999). The process of reading-to-write used by a skilled Spanish-as-a-foreign-language student: A case study. *Foreign Language Annals*, *32*(1), 45-58.

Ruiz-Funes, M. (2001). Task representation in foreign language reading-to-write. *Foreign Language Annals*, *34*(3), 226-234.

Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strengths and weaknesses: Assessing performance on an integrated writing task. *Language Assessment Quarterly*, *10*, 73-95.

Shetzer, H., & Warschauer, M. (2000). An electronic literacy approach to network-based language teaching. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 171-185). New York: Cambridge University Press.

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication*, *21*(2), 171-200.

Shi, L. (2010). Textual appropriation and citing behaviors of university undergraduates. *Applied Linguistics*, *31*(1), 1-24.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19-37). Charlotte, NC: Information Age Publishing, Inc.

Smoke, T. (2005). *A writer's workbook: A writing text with readings* (4th ed.). Cambridge: Cambridge University Press.

Sorapure, M., Inglesby, P., & Yatchisin, G. (1998). Web literacy: Challenges and opportunities for research in a new medium. *Computers and Composition*, *15*, 409-424.

Stapleton, P. (2001). Critical thinking in Japanese L2 writing: Implications about content familiarity and assumptions. *Written Communication*, *18*(4), 506-548.

Stapleton, P. (2003). Assessing the quality and bias of web-based sources: Implications for academic writing. *Journal of English for Academic Purposes*, *2*(3), 227-243.

Stapleton, P. (2005a). Evaluating web-sources: Internet literacy and L2 academic writing. *ELT Journal*, *59*(2), 135-143.

Stapleton, P. (2005b). Using the web as a research source: Implications for L2 academic writing. *Modern Language Journal*, *89*(2), 177-189.

Stapleton, P. (2012). Gauging the effectiveness of anti-plagiarism software: An empirical study of second language graduate writers. *Journal of English for Academic Purposes*, *11*, 125-133.

Stapleton, P., Helms-Park, R., & Radia, P. (2006). The Web as a source of unconventional research materials in second language academic writing. *The Internet and Higher Education*, *9*(1), 63-75.

Swales, J. M., & Lindemann, S. (2002). Teaching the literature review to international graduate students. In A. M. Johns (Ed.), *Genre in the classroom: Multiple perspectives* (pp. 105-119). Mahwah, NJ: Lawrence Erlbaum Associates.

Tannenbaum, R. J., & Katz, I. R. (2008). *Setting standards on the Core and Advanced iSkills™ assessments* (ETS RM-08-04). Princeton, NJ: Educational Testing Service.

Thompson, C., Morton, J., & Storch, N. (2013). Where from, who, why and how? A study of the use of sources by first year L2 university students. *Journal of English for Academic Purposes*, *12*, 99-109.

Watanabe, Y. (2001). Read-to-write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions. (Unpublished doctoral dissertation). University of Hawaii.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(1), 27-55.

Weigle, S. C., & Jensen, L. (1997). Assessment issues for content-based instruction. In M. A. Snow & D. Brinton (Eds.), *The content-based classroom: Perspectives on integrating language and content* (pp. 201-212). White Plains, NY: Addison Wesley Longman.

Weigle, S. C., & Montee, M. (2011, October). Textual borrowing and rater perceptions in integrated writing tasks. Unpublished paper presented at ECOLT 2011, Washington, DC.

Weigle, S., & Parker, K. (2011, March). Source text borrowing in an integrated reading/writing assessment. Unpublished paper presented at AAAL 2011, Chicago, IL.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Hampshire, UK: Palgrave Macmillan.

Wette, R. (2010). Evaluating student learning in a university-level EAP unit on writing using sources. *Journal of Second Language Writing*, *19*(3), 158-177.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal*, *46*(4), 1060-1106.

Wilson, K. (1997). "Wording" it up: Plagiarism in the interdiscourse of international students. In *Advancing international perspectives* (pp. 763-770), Proceedings of the International Conference of the Higher Education Research and Development Society of Australasia.

Wolfersberger, M. (2013). Refining the construct of classroom-based writing-from-readings assessment: The role of task representation. *Language Assessment Quarterly*, *10*, 49-72.

Wu, R.-J. R. (2013). Native and non-native students' interaction with a text-based prompt. *Assessing Writing*, *18*, 202-217.

Yang, H.-C. (2009). Exploring the complexity of second language writers' strategy use and performance on an integrated writing test through structural equation modeling and qualitative approaches. (Unpublished doctoral dissertation). The University of Texas at Austin.

Yang, H.-C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, *46*(1), 80-103.

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing*, *25*(4), 521-551.

Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing*, *14*, 116-137.

Yu, G. (2013). The use of summarization tasks: Some lexical and conceptual analyses. *Language Assessment Quarterly*, *10*, 96-109.

Zhang, W. (2003). *Doing English digital: An assessment model for a new college English curriculum in China*. (Unpublished doctoral dissertation). Columbia University.

Zhang, W. (2005). Digital literacy assessment: A model for a new college English curriculum. *Foreign Language Teaching and Research*, *38*(2), 115-121.

APPENDIX A

PREVIOUS PROMPTS FOR FINAL ESSAY

Possible prompts for final essay

Choose to give either an in-class persuasive or response essay during the final exam time. Make sure students are aware of which type of essay they will be writing during this time and corresponding evaluation criteria. If they are writing a response essay, make sure to tell students which essay they will be responding to ahead of time and tell them to bring their books/copy of the essay to class so they can refer to it if needed. Length will likely be between 1-3 pages.

Persuasive Essay

1. Which would you choose: a high-paying job with long hours that would give you little time with family and friends or a lower-paying job with shorter hours that would give you more time with family and friends? Explain your choice, using specific reasons and details.
2. If you were an employer, what kind of worker would you rather hire: an inexperienced worker at a lower salary, or an experienced worker at a higher salary? Use specific reasons and details to support your answer.

Evaluation Criteria

Material: (35 points)
-includes a brief reformulation of the scenario in the first paragraph
-clearly takes a position on a topic and includes ideas or points followed by examples that are in support of that position
-concludes with a restatement of the position

Organization: (30 points)
-material is organized appropriately to allow readers to clearly understand the author's main point and how information is each paragraph supports the thesis

Expression: (20 points)
-uses appropriate vocabulary—including transitional devices—and sentence structure to convey meaning clearly and maintain a reader's interest

Correctness: (15 points)
-uses appropriate word choice, sentence structure, punctuation, and spelling with few grammatical errors

Response Essay

1. Write a response to an essay that is in the book, e.g., "Village is more global, language is more vital" or "Cultural Identity vs. Ethnic Fashions."
2. Write a response to another essay of your choosing that is not in the book.

Evaluation Criteria

Material: (35 points)
-includes a brief summary of the text in the first paragraph
-focuses response on a specific passage or central point in the text
-responds by agreeing or disagreeing, explaining how it is relevant to their own experience, or expands on it by giving more information
-contains a conclusion summarizing or reinforcing the main point of the response

Organization: (30 points)
-material is organized appropriately to allow readers to clearly understand the author's main point and how information in each paragraph supports the thesis

Expression: (20 points)
-uses appropriate vocabulary—including transitional devices—and sentence structure to convey meaning clearly and maintain a reader's interest

Correctness: (15 points)
-uses appropriate word choice, sentence structure, punctuation, and spelling with few grammatical errors

APPENDIX B

TEST TASK SPECIFICATION

TITLE: Web-search-permitted and web-source-based integrated writing test
TEST USE: English 101C final exam
LEVEL: English 101C students (intermediate to high-intermediate learners of English)

General Description (GD)

General Objectives

Students will write a multiparagraph essay on an assigned topic using one or more sources that they have searched for on the internet and with the help of online language help options.

Specific Objectives

Students will demonstrate in writing their ability to express their ideas, thoughts, and/or opinions within paragraphs while writing a web-source-based argumentative essay on an assigned topic. In so doing, students will:

- Address the writing task
- Search the web for credible and reliable sources on the topic
- Present clear organization and development of paragraphs
- Use details and/or examples (students' own plus at least one web source) to support a thesis or illustrate an idea
- Display facility in the use of language
- Exhibit grammatical accuracy and correctness of citation, spelling, and punctuation

Prompt Attributes (PA)

Students will be asked to write an argumentative essay in response to a specific topic. Students will be required to search for and use information from at least one web source in their essay. Requirements for the selection of a topic include the following characteristics:

- A topic that is meaningful, relevant, and motivating to written communication at the college-level
- A topic that yields enough reliable and credible sources in a web search

Instructions:

1. You will have 2 hours to write an essay that answers the essay question printed below.
2. Your essay should include an introduction, body, and conclusion.
3. Your thesis and main ideas must be supported by information from one or more credible internet sources (citing the sources correctly in-text and in a references list) as well as your own insights and experience.
4. You may take notes or plan your essay in the blank Word document. You may also wish to type your essay in Word first and then copy/paste the completed essay into the text box below.
5. You are allowed to use online help options, such as dictionaries, thesauruses, grammar checkers, or citation-producing websites.
6. Aim to write at least 300 words.

7. Your essay will be evaluated on material, organization, expression, correctness, and use of internet sources.

Response Attributes (RA)
Students will write a web-source-based argumentative essay on the assigned topic. They will turn in the essays at the end of the two-hour exam period for assessment based on:

a. Material
b. Organization
c. Expression and word choice
d. Correctness of grammar, spelling, and punctuation
e. Choice of web sources and integration of information from sources

In this way, students can reflect on their own achievement as writers having taken English 101C.

Sample Item (SI)
Topic: Video games in education
Essay question: Should video games be used in elementary schools?

Specification Supplement (SS)
Rating rubric (Appendix K)

APPENDIX C

TEST PROMPT

Topic: Video games in education

Instructions:

1. You will have 2 hours to write an essay that answers the essay question printed below.
2. Your essay should include an introduction, body, and conclusion.
3. Your thesis and main ideas must be supported by information from one or more credible internet sources (citing the sources correctly in-text and in a references list) as well as your own insights and experience.
4. You may take notes or plan your essay in the blank Word document. You may also wish to type your essay in Word first and then copy/paste the completed essay into the text box below.
5. You are allowed to use online help options, such as dictionaries, thesauruses, grammar checkers, or citation-producing websites.
6. Aim to write at least 300 words.
7. Your essay will be evaluated on material, organization, expression, correctness, and use of internet sources.

Essay question: Should video games be used in elementary schools?

APPENDIX D

POST-TEST QUESTIONNAIRE

Post-test Questionnaire

The purpose of this questionnaire is to understand your perceptions of the writing test that you have just taken as well as your prior experiences of using the internet and writing in English. All possible measures will be taken to ensure the confidentiality of your personal information.

Please answer the following questions.

A. Perceptions of Test

1. Circle your rating of each statement.

	strongly disagree----------					----------strongly agree
The writing test was interesting.	1	2	3	4	5	6
The writing test was challenging.	1	2	3	4	5	6
The directions in the prompt were clear.	1	2	3	4	5	6
The amount of time given (2 hours) was adequate.	1	2	3	4	5	6
A writing test like this should be used in academic writing courses as a diagnostic test at the beginning of the semester or a final test at the end of the semester.	1	2	3	4	5	6
A writing test like this should be used in placement tests for new international students at universities in the US.	1	2	3	4	5	6
A writing test like this should be used in official English tests like the TOEFL or IELTS.	1	2	3	4	5	6

2. What did you like about the writing test?

3. What did you dislike about the writing test?

2. Prior Experience with the Internet

How often do you use the internet for the following activities? (Check all that apply.)

	On occasion	Once a week	Several times a week	Everyday
Web search (e.g., Google, Bing, Yahoo)				
Social networking services (e.g., Facebook)				
Reading online newspaper or magazine articles				
Reading online academic journal articles				
Email				
Reading blogs, message boards, or forums				
Writing on a blog, message board, or forum				

3. Prior Experience with Source-based Writing

Choose your rating of each statement (1=strongly disagree, 6=strongly agree).

I have written an essay or research paper <u>in my first language</u> based on source materials such as journal articles, books, newspaper articles, magazine articles, and webpages.	1	2	3	4	5	6
I have written an essay or research paper <u>in English</u> based on source materials such as journal articles, books, newspaper articles, magazine articles, and webpages.	1	2	3	4	5	6
I refer to printed materials (e.g., hard copies of books, journals, magazines, or newspapers) when I am writing an essay or research paper.	1	2	3	4	5	6
I search for source materials on the internet when I am writing an essay or research paper.	1	2	3	4	5	6
I have learned how to cite outside sources in my essays or research papers.	1	2	3	4	5	6
I am confident about citing outside sources in my essays or research papers.	1	2	3	4	5	6

4. Personal Background

1. Name: _____
2. Major: _____
3. Age: _____
4. Gender: ____ Female ____ Male
5. First language: _____
6. How many years have you been studying English? _____ years
7. How long have you been in the U.S.? _____ years _____ months
8. Your most recent English test scores: PBT TOEFL: _____
   CBT TOEFL: _____
   iBT TOEFL: _____
   IELTS: _____
   Other (Please specify:_____): _____

Please check the times when you are available if you would like to participate in a 20-minute interview to talk about your thoughts and experiences of the writing test.

		Check boxes below. ↓
Thursday, Dec 15	Morning (9am-12pm)	
	Early afternoon (12-3pm)	
	Late afternoon (3-6pm)	
Friday, Dec 16	Morning (9am-12pm)	
	Early afternoon (12-3pm)	
	Late afternoon (3-6pm)	
Saturday, Dec 17	Morning (9am-12pm)	
	Early afternoon (12-3pm)	
	Late afternoon (3-6pm)	

APPENDIX E

SEMI-STRUCTURED INTERVIEW PROTOCOL FOR POST-TEST INTERVIEWS WITH STUDENTS

1. Can you describe your experience of taking the writing test?
2. How did you feel about the writing test?
3. How did you use internet sources in the test? What strategies did you try to employ when searching for sources on the internet?
4. Are there particular websites that you often refer to when writing essays for English or other classes?
5. What were some things you liked about the writing test?
6. What were some things that you did not like about the writing test?
7. What were some problems you experienced during the test?
8. How do you think the test-taking experience can be improved?
9. What suggestions do you have to improve the writing test?
10. What kind of writing tests have you taken before?
11. Was this test similar to any writing you have done before?
12. Did the topic affect your writing?

APPENDIX F

FOLLOW-UP QUESTIONNAIRE FOR STUDENTS

1. Name: _____

2. Semester in which you took English 101C (circle one): Fall 2011 Spring 2012

3. Scores on writing sections of English proficiency tests (if known or available):
   a. iBT TOEFL _____/30
   b. IELTS _____/9.0
   c. PBT/CBT TOEFL (TWE) _____/6

4. Grades on writing assignments completed in courses beyond English 101C:

   a. English 150: _____

   b. English 250: _____

   c. Other courses: _____

5. Please take a look at the attached grading rubric which was developed to grade the final essay exam for English 101C.
   a. How <u>useful</u> are the score descriptors for understanding strengths and weaknesses in writing?

   (not useful at all) 1 2 3 4 5 6 (very useful)

   b. How <u>clear</u> are the score descriptors?

   (not clear at all) 1 2 3 4 5 6 (very clear)

   c. How <u>easy to interpret</u> are the score descriptors?

   (not easy at all) 1 2 3 4 5 6 (very easy)

6. Did you think that the instruction you received in English 101C was adequate for taking the final essay exam?

   (not adequate at all) 1 2 3 4 5 6 (very adequate)

   Why do you think so?

7. When you learned about the requirements and details of the final essay exam in English 101C, how did you prepare for the test?

APPENDIX G

SEMI-STRUCTURED INTERVIEW PROTOCOL FOR FOLLOW-UP INTERVIEWS WITH STUDENTS

1. How did you do or how are you doing in English 150, English 250, or other courses that require writing assignments? Can you describe some examples of writing assignments you have completed?
2. How have you performed or are you performing on source-based assignments in English 150, English 250, or other courses? Can you describe some experiences of using sources in writing assignments?
3. Please take a look at the attached grading rubric which was developed to grade the final essay exam for English 101C. You will have seen this before in the survey.
    a. How useful are the score descriptors for understanding strengths and weaknesses in writing? 1 (not useful at all) – 6 (very useful)
    b. How clear are the score descriptors? 1 (not clear at all) – 6 (very clear)
    c. How easy to interpret are the score descriptors? 1 (not easy at all) – 6 (very easy)

    Can you elaborate on the reasons for your choices? How would you use the information from this rubric if you were to have received one after the final exam?
4. Did you think that the instruction you received in English 101C was adequate for taking the final essay exam? 1 (not adequate at all) – 6 (very adequate) Why do you think so? Do you think that students were given equal opportunity to prepare for the exam?
5. When you learned about the requirements and details of the final essay exam, how did you prepare for the test?
6. How helpful was the instruction you received in English 101C for completing source-based writing assignments in English 150, English 250, or other courses? How did you use what you learned from English 101C in other courses?
7. What did you get out of the experience of taking the final essay exam in English 101C?
8. Please take a look at the exam prompt. How clear are the directions? How clear is the requirement to use at least one outside source?
9. Please take a look at your essay. Could you describe how you used sources?
10. Was English 101C the first time you learned to cite sources? Was it the first time you had to use sources in writing?
11. Have you had to use and cite sources beyond English 101C? How did you do it? How well did you do?
12. How important is it to know how to use and cite sources in writing?
13. Do you think “using and citing sources” should be taught in English 101C?

APPENDIX H

SEMI-STRUCTURED INTERVIEW PROTOCOLS FOR EXPERTS AND STAKEHOLDERS

Instructors of English 101C

1. What writing assignments and tasks are required in the course that you teach?
2. What language skills, knowledge, and abilities are taught in your course? What skills, knowledge, and abilities are important for success in your course?
3. Do you teach students how to use (web) sources in their writing? If so, how do you achieve that?
4. How important do you think source-based writing ability is for success at the university?
5. Do you introduce online or offline resources and help options for writing to students? If so, what are they?
6. How would you design a task that tests source-based academic writing?
7. What criteria or aspects should appear in a rubric that scores essays from a test of source-based academic writing?
8. Please take a look at this test specification. Do you have any suggestions for improvement or modification?
9. Based on this test specification, would you be able to create parallel tasks?
10. To what extent does the test task reflect the instructional tasks that are used in your course?
11. Do you think the test task requires important skills taught in your course? How representative of the domain of source-based writing in college courses is the test task?
12. What do you think about the wording of the prompt? Do you have suggestions for improvement or modification?
13. To what extent does the rating rubric reflect the rubrics that are used in your course?
14. Take a look at these score descriptors that will be provided to test takers along with their test scores. How useful, clear, and interpretable do you find the score descriptors from the perspective of the instructor? How about from the students' perspective?
15. How would you feel about using this test as a final exam in English 101C?
16. If you were to use this test as a final exam in your class, what might be some potential washback effects of using this test on your teaching and test preparation?

Instructors of English 150

1. What writing assignments and tasks are required in the course that you teach?
2. What language skills, knowledge, abilities, and processes are taught in your course? What skills, knowledge, abilities, and processes are important for success in your course?
3. Do you teach or require students how to use (web) sources in their writing? If so, how do you achieve that?
4. How important do you think source-based writing ability is for success at the university?
5. Do you introduce online or offline resources and help options for writing to students? If so, what are they?
6. Please take a look at this test specification. Please let me know if there are any parts that you do not understand. If such a test was given at the end of English 101C as a final exam and a student performed well on the test, would you say that the student is prepared for the demands and requirements of writing in your course?
7. Take a look at these score descriptors that will be provided to test takers along with their test scores. How useful, clear, and interpretable do you find the score descriptors? Do the descriptors give you a clear picture of an English 101C student's source-based academic writing ability?

APPENDIX I

PILOT RATING RUBRIC

Level 4 (A)	• Contains an Intro, Body and Conclusion • Clear thesis statement, appropriately placed • Good development of thesis; logical sequencing; reasonable use of transitions • Paragraphs are fairly cohesive • Good synthesis of ideas • Summary of source content may contain minor inaccuracies, but good understanding is indicated; effective, skillful paraphrase • Sources are cited, though possibly inaccurately • May contain minor grammatical/lexical errors, but meaning is clear • Strong linguistic expression exhibiting academic vocabulary, sentence variety, and complexity
Level 3 (B)	• Length is sufficient for full expression of ideas • Writes on topic • Elements of essay organization are clearly present, though they may be flawed • Attempt to advance a main idea; presence of thesis statement • Flows somewhat smoothly • Some development and elaboration of ideas; evidence of logical sequencing; transitions may show some inaccuracies • Paragraph structure generally mastered, generally cohesive • Attempts to use sources to advance the thesis; evidence of some synthesis of ideas • Use of sources demonstrates basic understanding • Covert plagiarism; attempted summary and paraphrase; may contain isolated instances of direct copying; may not cite sources, OR may cite them incorrectly • Moderately successful paraphrase in terms of smoothness • Some grammatical/lexical errors; meaning may be occasionally obscured, but essay is still comprehensible • Inconsistent evidence of some sophistication in sentence variety and complexity
Level 2 (C)	• Length may be insufficient to evaluate; may be off-topic • Elements of essay organization (Intro, Body and Conclusion) may be attempted, but are simplistic and ineffective • Essay may lack a central controlling idea (no thesis statement, OR thesis statement is flawed) • Essay does not flow smoothly; ideas are difficult to follow • Development of ideas is insufficient; examples may be inappropriate; logical sequencing may be flawed or incomplete • Paragraph structure not mastered; lack of main idea (topic sentence), focus, and cohesion • Summarizes/restates sources rather than using them to support ideas • May lack synthesis of ideas (of sources or of sources and student's own ideas) • May indicate misunderstanding of source material • Attempts to paraphrase are generally unskillful and inaccurate • Some overt plagiarism

	• Grammatical and lexical errors impede understanding; awkwardness of expression; general inaccuracy of word forms • Little sophistication in vocabulary and linguistic expression; little sentence variety; sentence complexity not mastered
Level 1 (D)	• Length insufficient to evaluate • No organization of ideas; no cohesion; like a free writing • Content marked by inaccuracies of source information, OR content is completely off-topic, OR majority of essay is copied • Grammatical and lexical errors are severe; no complexity; even simple sentences are flawed

Adapted from Benchmarks for EPT composition scoring: Graduate essays (Revised 01/05; Ann Spear), UIUC

APPENDIX J

PREVIOUS RATING RUBRIC

Rating Rubric for Final Exam: Web-Source-Based Argumentative Essay

Criteria	Level 4 (A) Excellent	Level 3 (B) Good	Level 2 (C) Satisfactory	Level 1 (D) Needs Work
Material: Fully develops an argument for a position using interesting and appropriate examples and supporting details; must include information from at least one (credible) internet source in the form of a quote, paraphrase, or summary	**30** • Length is sufficient for full expression of ideas • Writes on topic • Good development of thesis • Good synthesis of ideas • Summary of source content may contain minor inaccuracies, but good understanding is indicated; effective, skillful paraphrase	**26** • Length is sufficient for full expression of ideas • Writes on topic • Some development and elaboration of ideas • Attempts to use sources to advance the thesis; evidence of some synthesis of ideas • Use of sources demonstrates basic understanding • Covert plagiarism; attempted summary and paraphrase; may contain isolated instances of direct copying • Moderately successful paraphrase in terms of smoothness	**22** • Length may be insufficient to evaluate; may be off-topic • Development of ideas is insufficient; examples may be inappropriate • Summarizes/restates sources rather than using them to support ideas • May lack synthesis of ideas (of sources or of sources and student's own ideas) • May indicate misunderstanding of source material • Attempts to paraphrase are generally unskillful and inaccurate • Some overt plagiarism	**18** • Length insufficient to evaluate • Content marked by inaccuracies of source information, OR content is completely off-topic, OR majority of essay is copied
Organization: Material is organized appropriately to allow readers to clearly understand the author's stance and how information in each paragraph supports that position; includes an introduction with a clear thesis; provides coherent transitions from one idea to another; well-developed paragraphs (topic sentence, unity, cohesion, etc.); a clear conclusion	**30** • Contains an Intro, Body and Conclusion • Clear thesis statement, appropriately placed • Logical sequencing; reasonable use of transitions • Paragraphs are fairly structured, unified, and cohesive	**26** • Elements of essay organization are clearly present, though they may be flawed • Attempt to advance a main idea; presence of thesis statement • Flows somewhat smoothly • Evidence of logical sequencing; transitions may show some inaccuracies • Paragraph structure generally mastered, generally	**22** • Elements of essay organization (Intro, Body and Conclusion) may be attempted, but are simplistic and ineffective • Essay may lack a central controlling idea (no thesis statement, OR thesis statement is flawed) • Essay does not flow smoothly; ideas are difficult to follow • Logical sequencing may be	**18** • No organization of ideas; no cohesion; like a free writing

		cohesive	flawed or incomplete • Paragraph structure not mastered; lack of main idea (topic sentence), focus, and cohesion	
Expression (word choice, fluency): Uses appropriate vocabulary and sentence types to convey meaning clearly and maintain a reader's interest	**20** • Strong linguistic expression exhibiting academic vocabulary, sentence variety, and complexity	**17** • Inconsistent evidence of some sophistication in vocabulary, sentence variety, and complexity	**14** • Little sophistication in vocabulary and linguistic expression; awkwardness of expression; little sentence variety; sentence complexity not mastered	**11** • No complexity
Correctness: Uses appropriate word forms, sentence structure, punctuation, and spelling with few grammatical errors; must use some form of in-text citation and references list	**20** • May contain minor grammatical/lexical errors, but meaning is clear • Sources are cited, though possibly inaccurately	**17** • Some grammatical/lexical errors; meaning may be occasionally obscured, but essay is still comprehensible • May not cite sources, OR may cite them incorrectly	**14** • Grammatical and lexical errors impede understanding; general inaccuracy of word forms • May not cite sources, OR may cite them incorrectly	**11** • Grammatical and lexical errors are severe; even simple sentences are flawed • May not cite sources, OR may cite them incorrectly
Total				**/100**

Adapted from Benchmarks for EPT composition scoring: Graduate essays (Revised 01/05; Ann Spear), UIUC

APPENDIX K

REVISED RATING RUBRIC

Rating Rubric for Final Exam: Web-Source-Based Argumentative Essay

Criteria	Level 4 (A) Excellent	Level 3 (B) Good	Level 2 (C) Satisfactory	Level 1 (D) Needs Work
Material: Fully develops an argument for a position using interesting and appropriate examples and supporting details; must include information from at least one (credible) internet source in the form of a quote, paraphrase, or summary	**30** • Length of essay is sufficient for full expression of ideas (at least 300 words) • Writes on topic • Full development of thesis through interesting and appropriate supporting details, examples, and information from sources • Good synthesis of ideas • Effective and skillful integration of information from sources, showing good understanding of source content	**26** • Length of essay is sufficient for full expression of ideas (at least 300 words) • Writes on topic • Some development and elaboration of ideas • Attempts to use sources to advance the thesis; evidence of some synthesis of ideas • Use of sources demonstrates basic understanding • Moderately successful integration of source content	**22** • Length of essay is sufficient for full expression of ideas (at least 300 words) • Development of ideas is insufficient; examples may be irrelevant or inappropriate • Summarizes/restates sources rather than using them to support ideas • May lack synthesis of ideas (of sources or of sources and student's own ideas) • Attempts to integrate source content are generally unskillful and inaccurate • May indicate misunderstanding of source material	**18** • Length of essay may be insufficient to evaluate • Content marked by inaccuracies of source information, OR content is completely off-topic, OR majority of essay is copied • Overt plagiarism
Organization: Material is organized appropriately to allow readers to clearly understand the author's stance and how information in each paragraph supports that position; includes an introduction with a clear thesis; provides coherent transitions from one idea to another; well-developed paragraphs (topic sentence, unity, cohesion, etc.); a clear conclusion	**30** • Contains a clear introduction, body and conclusion • Clear thesis statement, appropriately placed • Logical sequencing; effective use of transitions • Paragraphs are well structured, unified, and cohesive	**26** • Elements of essay organization are clearly present, though there may be minor problems • Attempt to advance a main idea; presence of thesis statement • Flows somewhat smoothly • Evidence of logical sequencing; transitions may show some inaccuracies • Paragraph structure generally mastered, generally	**22** • Elements of essay organization (introduction, body and conclusion) may be attempted, but are simplistic and ineffective • Essay may lack a central controlling idea (no thesis statement, OR thesis statement is flawed) • Essay does not flow smoothly; ideas are difficult to follow • Logical sequencing may be	**18** • No organization of ideas; no cohesion; like a free writing

		cohesive	flawed or incomplete • Paragraph structure not mastered; lack of main idea (topic sentence), focus, and cohesion	
Expression (word choice, fluency): Uses appropriate vocabulary and sentence types to convey meaning clearly and maintain a reader's interest	**20** • Strong linguistic expression exhibiting academic vocabulary, sentence variety, and complexity	**17** • Evidence of some sophistication in vocabulary, sentence variety, and complexity	**14** • Little sophistication in vocabulary and linguistic expression; awkwardness of expression; little sentence variety; sentence complexity not mastered	**11** • No complexity
Correctness: Uses appropriate word forms, sentence structure, punctuation, and spelling with few grammatical errors; uses some form of in-text citation and references list	**20** • May contain a few minor grammatical/lexical/ mechanical errors, but meaning is clear • Sources are cited	**17** • Some grammatical/lexical/ mechanical errors; meaning may be occasionally obscured, but essay is still comprehensible • May not cite sources, OR may cite them incorrectly	**14** • Grammatical, lexical, and mechanical errors impede understanding; general inaccuracy of word forms • May not cite sources, OR may cite them incorrectly	**11** • Grammatical, lexical, and mechanical errors are severe; even simple sentences are flawed • May not cite sources, OR may cite them incorrectly
Total				**/100**

APPENDIX L

PILOT CODING SCHEME FOR SCREEN RECORDING DATA

Search engine
-Google (key word(s) typed in)
-iciba.com (key word(s) typed in)
-Baidu (key word(s) typed in)

Dictionary
-Dict.cn (word(s) typed in)
-Dictionary.com (word(s) typed in)
-Longman (word(s) typed in)
-Merriam-Webster (word(s) typed in)
-The Free Dictionary (word(s) typed in)

Criterion®
-Spell checker

Google Translate (key word(s) typed in)

Read
-prompt

Click from search results
-1st link/article (web page title/URL)
-2nd link/article (web page title/URL)
-3rd link/article (web page title/URL)
-xth link/article (web page title/URL)

Read
-1st link/article (web page title/URL)
-2nd link/article (web page title/URL)
-3rd link/article (web page title/URL)
-xth link/article (web page title/URL)

Draft/write (text box/Word document)
-outline/skeleton
-introduction
-body
-conclusion
-use source (copy/paste, quote, paraphrase)
-cite (in-text citation, other attribution)
-references list/works cited
-revise/edit
-word count

APPENDIX M

FINAL CODING SCHEME FOR SCREEN RECORDING DATA

Read prompt

Search Search engine name Key word(s)
Read search results page
Read 2nd search results page
Click on xth link Web page title Source type/URL
Read xth link Web page title Source type/URL

Look up dictionary Dictionary name Word(s) looked up
Read results page
Read definitions page

Translate Translator name Language 1 → Language 2 Word(s) typed in

Write (in text box/in Word doc)
- Title
- Introduction
- Body paragraph #
- Conclusion

Highlight and copy (word/phrase/sentence/paragraph)
Paste (word/phrase/sentence/paragraph) into (text box/Word doc)
Add parenthetical citation

Create references list
Add references list entry
Use (Citation Machine) to create references list entry Going back and forth from xth link

Read essay
Think

Proofread/edit

Right click on X and consider options
Right click on X and choose Y (Word/Mac)'s (spell/grammar) checker (red/green) squiggle
Make small change to remove (red/green) squiggle

Word count
Look for Word count

Adjust windows
Close windows

APPENDIX N

RATING GUIDE AND BENCHMARK ESSAYS

Rating guide

Dear rater,

Thank you so much for your valuable time and help! Please follow these steps:

1. Look over the essay rating rubric on pages 2-3 to become familiar with the four criteria as well as the score descriptors for the four score levels under each criterion.

2. Skim the four sample benchmark essays (pages 4-7), one for each of the four overall levels (A to D), to get a general idea of the expectations of performance at each level.

Read and rate the essays in the PDF document titled “Essays_rater#.” Record the scores in the Excel spreadsheet titled “Rating_sheet.” Provide four sub-scores for each essay. The total score should appear automatically. If possible, please also note the time that it took to rate each essay to the nearest minute.

(A)

Nowadays, with the developing of the technology, computers are used in many areas. Even video games have been used in elementary schools. Some people think that it not a good idea to use video games to help teacher to teach. They think that it wastes time and children do it just for fun, not for learning. But in my opinion, using video games to help teacher to teach is a great way for both teachers and students.

Firstly, video games help kids to exercise. “A recent study from the University of Oklahoma showed that active video games like Wii boxing or Dance Dance Revolution get kids as active as if they were taking a walk. “(1*) Sometime, kids will fear or don’t like something they haven’t known.Teachers can use video games to help them get familiar with the sports. If the kids like it, they will try to do the sport, if they don’t like some sports, the video games can also help them to do the exercise. Therefore, video games are a good way to help kids in exercising.

Secondly, video games could help kids to exercise for intelligence. This will release the tremendous creative powers of the kids. The strategic thinking and problem solving during video games makes them good learning machines. “For example, the hidden treasure is in the castle. They engage in an action by hunting for the treasure. Gamers discover if their hypothesis is true or false when they search the castle. If they don't find the treasure, they revise their hypothesis the next time they play. Video games are goal-driven experiences, says Gee” (2*) Therefore, video games are a wonderful way to help kids in fundamental learning.”

Finally, video games are not only help the kids but also help the teacher to make the class more interesting. We all know that history is very boring for kids, but it will be different by using History-based computer games to teach. “History-based computer games like Civilization and Age of Empires as well as life-simulating titles like Sims 2 and the city planning game SimCity are creeping into the curriculum of many classrooms, Squire said.”(3*) Therefore, video games are an excellent way to teach boring classes.

From above information, we can clearly see that video games are a good method to use for learning and teaching in elementary schools. Kids will do body building and brain training by it and teacher will use it to make class more interesting.

Resources:

1.http://www.pediatricsafety.net/2010/06/wii-helps-special-needs-kids-get-exercise/

2.http://parenting.kaboose.com/behavior/video-games-smart.html

3.http://www.sfgate.com/cgi-bin/article.cgi?f=/c/a/2006/02/20/CLASSROOM.TMP

(B)

Video Games VS Elementary Schools

With the development of new technologies, students find a new way to forget pressure they get from study and daily life by driving themselves into a virtual world called video games. Should video games be used in elementary school can be considered both the advantages and disadvantages.

Let the bad side be shown at first. When video games help students to relax, they make students fall into and become addicted easily at the same time. According to a study in 2008 of 1,178 children in the US, almost 9 percent of child gamers were pathologically or clinically "addicted" to playing video games. However, an elementary school is an institution where children receive the first stage of compulsory education, which means this part of students are so young to this world that they are weak at distinguishing new things and control themselves. After these children get contact to video games, it is possible that they will be addicted in it since childhood and bad prepared for the following education.

However, what if elementary schools help their students to develop the good habits about video games by regular management and correct guide? Actually children have their rights to have a free and relaxed childhood. If kids learn to manage time of playing video games and be prevented from addiction from the very beginning, it is obvious that tings will be different. Video games sometimes can even be a tool that used to motivate elementary students. In Huntsville, Alabama students are playing video games during school time. The program tracks the progress of students who participate and the teacher can check to determine if students are gaining in knowledge.

As an old saying told us, “ Things have two sides.” From their program, one thing should be mentioned is that the point of this problem is not should we use or not, but how to use.

Work cited:

1. http://en.wikipedia.org/wiki/Elementary_school

2.http://www.diyfather.com/content/Interesting_Statistics_About_Video_Games

3.http://elementaryschools.org/blog/elementary-school-controversy/video-games-motivate-elementary-students/

(C)

Are video games educational? Alone with the development of video games, they are not only a entertainment but also use as a education equipment recently. Some experts named this situation with edutainment, some of them support use video games but some don't. In my point of view it is good to use video games to teach in elementary schools.

Currently, the biggest problem of education is students are not interest in knowledge learning they may just like playing games. To use video games can definitely interested students in school education and also make them concentrated in class. A good video game always have challenges elements of surprise, between different level there are always excitememt regret and risk. And According to the research (by Jayel Gibson education.com) video games provide a new type of information and it is certainly more interesting than the traditional text and graphic. What's more it is also easy to memorize. As we all know, graphic is easier to be memorized than text and games can reward even better memory. During playing games, to win the games players should also matching pairs, answering question or problem solving. These points are also proved in a another research ("school uses video games to teach thinking skill" by Heather Chaplian). The instructor claimed that during educational gaming students learn how to solve problems, how to communicate, how to use date how to predict the coming difficulties. Experts call this the system thinking, the system thing is a important skill for students' future when they enter society. The students who involved in the video games teaching also give positive comments and they believe their school education.

Currently, all these experiments are done by single schools and it haven't spread. however I believe that as long as video games show its benefits in education. The method would exploded.

(D)
I think video games be not used in elementary school.Because elementary school children not are adult. some video games have more violent. The child is in control of the violence and experience the violence in his own eyes, then transmit some wrong information in his mind.

To much video games palying make children socially isolated, then less time to do homework, sport, there are effect on some children health.Children need more time in outside.Video games can stop children developing a proper imagination.When children palying online,they can pick some bad language and behavior from other people vulnerable to online danger.Children not have self-discipline,they don't know what is bad for him.

Some video games teach kids the wrong values,violent behavior,vengeance and aggression are rewarded.Negotiating and other nonviolent solutions are often not options,woman are often portrayed as weaker characters that are helpless or sexually provocative.

Also some children wallow in games, then don't like come to school or not spirit in class.In his mind only have games, be infatuated with games.In china,one child palying games about 20 hours, he forgot eating,sleeping,studying in Internet bar.When he finished games,he came to school,and set in the school windowsill,then he jump down.Because he was infatuated with games long time,his mind also in the game, he felt he can fly,he was strong in the game.So he was die.

In my mind video games looks like a devil.when you love that,you can't control yourself.Children can't palying video games when he is a kid. This is a good hobby.Speeding more time to studying or sporting. There are good for your health and mind.

APPENDIX O

FREQUENCY LIST OF WORDS INCLUDED IN SEARCH KEY WORDS

RANK	FREQUENCY	COVERAGE individual	cumulative	WORD
1.	100	17.51%	17.51%	VIDEO
2.	78	13.66%	31.17%	GAMES
3.	39	6.83%	38.00%	IN
4.	36	6.30%	44.30%	ELEMENTARY
5.	25	4.38%	48.68%	GAME
6.	23	4.03%	52.71%	USED
7.	22	3.85%	56.56%	BE
8.	20	3.50%	60.06%	EDUCATION
9.	20	3.50%	63.56%	SCHOOL
10.	17	2.98%	66.54%	SCHOOLS
11.	14	2.45%	68.99%	OF
12.	14	2.45%	71.44%	SHOULD
13.	12	2.10%	73.54%	FOR
14.	8	1.40%	74.94%	TEACHING
15.	7	1.23%	76.17%	THE
16.	7	1.23%	77.40%	TO
17.	6	1.05%	78.45%	CHILDREN
18.	5	0.88%	79.33%	EFFECTS
19.	5	0.88%	80.21%	HOW
20.	4	0.70%	80.91%	AND
21.	4	0.70%	81.61%	ARE
22.	4	0.70%	82.31%	VIOLENCE
23.	4	0.70%	83.01%	WHAT
24.	3	0.53%	83.54%	A
25.	3	0.53%	84.07%	CAN
26.	3	0.53%	84.60%	EDUCATIONAL
27.	3	0.53%	85.13%	GOOD
28.	3	0.53%	85.66%	IMPORTANT

29.	3	0.53%	86.19%	KIDS
30.	3	0.53%	86.72%	PLAY
31.	2	0.35%	87.07%	ADDICTION
32.	2	0.35%	87.42%	BENEFIT
33.	2	0.35%	87.77%	DISADVANTAGES
34.	2	0.35%	88.12%	ELEMETARY
35.	2	0.35%	88.47%	ESSAY
36.	2	0.35%	88.82%	EXAMPLES
37.	2	0.35%	89.17%	IS
38.	2	0.35%	89.52%	MAKE
39.	2	0.35%	89.87%	ON
40.	2	0.35%	90.22%	PLAYING
41.	2	0.35%	90.57%	RESEARCH
42.	2	0.35%	90.92%	STUDENTS
43.	2	0.35%	91.27%	TOOLS
44.	2	0.35%	91.62%	VEDIO
45.	2	0.35%	91.97%	YOU
46.	1	0.18%	92.15%	AGRESSION
47.	1	0.18%	92.33%	ANIMATION
48.	1	0.18%	92.51%	AS
49.	1	0.18%	92.69%	ATTENTION
50.	1	0.18%	92.87%	BAD
51.	1	0.18%	93.05%	BRAIN
52.	1	0.18%	93.23%	CHILDREN'S
53.	1	0.18%	93.41%	CITY
54.	1	0.18%	93.59%	DANGEROUS
55.	1	0.18%	93.77%	DEVELOPMENT
56.	1	0.18%	93.95%	DISADVANTAGE
57.	1	0.18%	94.13%	DRAWBACKS
58.	1	0.18%	94.31%	EA
59.	1	0.18%	94.49%	EDUTAINMENT
60.	1	0.18%	94.67%	ELEM

61.	1	0.18%	94.85%	ELEME
62.	1	0.18%	95.03%	ELEMEANTARY
63.	1	0.18%	95.21%	ELEMENTS
64.	1	0.18%	95.39%	EXERCISE
65.	1	0.18%	95.57%	GTA
66.	1	0.18%	95.75%	HEALTH
67.	1	0.18%	95.93%	HURT
68.	1	0.18%	96.11%	LEAD
69.	1	0.18%	96.29%	LIFE
70.	1	0.18%	96.47%	LIKE
71.	1	0.18%	96.65%	MANY
72.	1	0.18%	96.83%	MUCH
73.	1	0.18%	97.01%	NUMBERD
74.	1	0.18%	97.19%	OR
75.	1	0.18%	97.37%	OWN
76.	1	0.18%	97.55%	PEOPLE
77.	1	0.18%	97.73%	PRIMARY
78.	1	0.18%	97.91%	PROS
79.	1	0.18%	98.09%	SCHOO
80.	1	0.18%	98.27%	SCHOOLD
81.	1	0.18%	98.45%	SIM
82.	1	0.18%	98.63%	SMARTER
83.	1	0.18%	98.81%	STUDENT
84.	1	0.18%	98.99%	TEENAGE
85.	1	0.18%	99.17%	TOO
86.	1	0.18%	99.35%	USEFUL
87.	1	0.18%	99.53%	USING
88.	1	0.18%	99.71%	VEDEO
89.	1	0.18%	99.89%	VIDEOGAMES
90.	1	0.18%	100.00%	VIRTUAL
91.	1	0.18%	100.00%	WAY